DOCUMENT RESUME

ED 078 877 LI 004 412

AUTHOR	Fiorello, Marco R.
TITLE	Management and Design Tools for Document Retrieval Systems: A Method for Predicting Quantity Output.
INSTITUTION	Rand Corp., Santa Monica, Calif.
PUB DATE	Mar 73
NOTE	221p.;(153 References)
AVAILABLE FROM	The Rand Corporation, 1700 Main St., Santa Monica, Calif. 90406 ($5.00)
EDRS PRICE	MF-$0.65 HC-$9.87
DESCRIPTORS	*Design; Information Processing; *Information Retrieval; *Information Storage; *Information Systems; *Management; Search Strategies

ABSTRACT

The existing volume and increasing growth rate of documented information has resulted in numerous efforts to construct operational Document Storage and Retrieval Systems, as a practical solution to the demand for information storage and retrieval. Accompanying the surge to build more and better and bigger Document Retrieval Systems (DRSs), was the realization that there are few effective tools for the designers and managers of these systems. The tasks of design and management of DRSs requires tools and performance measures to aid in the selection of preferred options, and in the control over the fundamental processes of inquiry analysis, indexing, retrieval and system growth. A step toward the generation of operational tools to aid in the design and management tasks is presented in this report, by the development of a Retrieval Quantity (Rq) estimate. The Rq estimate is defined as a function of the inquiry form, search strategy and descriptor-document distribution, and can be used to predict the quantity output due to system growth, and aid in the tuning of the indexing and formal inquiry specification processes. (Author/NH)

ED 078877

AUG 16 1973

U.S. DEPARTMENT OF HEALTH,
EDUCATION & WELFARE
NATIONAL INSTITUTE OF
EDUCATION
THIS DOCUMENT HAS BEEN REPRODUCED EXACTLY AS RECEIVED FROM THE PERSON OR ORGANIZATION ORIGINATING IT. POINTS OF VIEW OR OPINIONS STATED DO NOT NECESSARILY REPRESENT OFFICIAL NATIONAL INSTITUTE OF EDUCATION POSITION OR POLICY

MANAGEMENT AND DESIGN TOOLS FOR DOCUMENT RETRIEVAL SYSTEMS:
A METHOD FOR PREDICTING QUANTITY OUTPUT

Marco R. Fiorello

March 1973

LI 004 412

P-4980

FILMED FROM BEST AVAILABLE COPY

Any views expressed in this paper are those of the authors. They should not be interpreted as reflecting the views of The Rand Corporation or the official opinion or policy of any of its governmental or private research sponsors. Papers are reproduced by The Rand Corporation as a courtesy to members of its staff.

ABSTRACT

The existing volume and increasing growth rate of documented information has resulted in numerous efforts to construct operational Document Storage and Retrieval Systems, as a practical solution to the demand for information storage and retrieval. Accompanying the surge to build more and better and bigger Document Retrieval Systems (DRSs), was the realization that there are few effective tools for the designers and managers of these systems. The tasks of design and management of DRSs requires tools and performance measures to aid in the selection of preferred options, and in the control over the fundamental processes of inquiry analysis, indexing, retrieval and system growth.

A step toward the generation of operational tools to aid in the design and management tasks is presented in this report, by the development of a Retrieval Quantity (R_q) estimate. The R_q estimate is defined as a function of the inquiry form, search strategy and descriptor-document distribution, and can be used to predict the quantity output of an inquiry, measure the impact on quantity output due to system growth, and aid in the tuning of the indexing and formal inquiry specification processes. The definition of the R_q measure is based on the identification of certain canonical forms which characterize the underlying principles of DRS indexing and retrieval. The R_q estimate was tested on an operational DRS, and demonstrated high prediction accuracy for a variety of typical inquiries. Though developed on a small DRS, the methodology for determining R_q appears to hold for a very wide range of system size, subject content and construction.

ACKNOWLEDGMENTS

I take this opportunity to acknowledge my debts to my colleagues at Rand and the Institute of Library Research at the University of California for the many fruitful conversations and assistance of various kinds that were so helpful in this task.

To my thesis committee of Dr. C. West Churchman, Dr. Bill Maron, and Dr. Franco Nicosia I am most indebted. They have been patient, understanding and always cooperative.

I am grateful to all these friends, and to each my fondest thanks.

LIST OF TABLES

Table	Title	Page
3.1	Retrieval Quality Performance Measures	47
5.1	Institute of Library Research Document Retrieval System Characteristics	71
5.2	Sample System Characteristics	71
5.3	Comparison of System Characteristics for: The Institute of Library Research, and Systems Investigated by Litofsky (90), Houston and Wall (68), A. D. Little (1, 2), Wall (143)	85
5.4	MEZ Parameters and Function Values	90
5.5	Comparison of γ and γ_1 Factors	102
5.6	Test Inquiries	142
5.7	Comparison of Actual Versus Estimated Quantity Output	144
5.8	Term-Term Co-occurrences Between Terms with Different Frequency of Use	151
5.9	Coefficiencies of Association Parameter -- α	154
5.10	Comparison of Outputs for Stage 1 and Stage 2 for the Test System	157

LIST OF FIGURES

Figure		Page
1.1	Estimate of File Size of Books and Periodicals in U.S. Colleges	4
1.2	Estimate of File Size of U.S. Public Libraries	5
1.3	Estimate of the Number of Technical Literature Abstracts Produced in the World	6
1.4	Growth of Journals and Abstract Journals	7
1.5	Literature Growth in Economics and Other Professional Fields	8
1.6	Information Storage and Retrieval Systems	10
2.1	Functional Description of Document Retrieval Systems	14
2.2	Hierarchical Systems	21
2.3	Facet Systems	22
2.4	Taxonomy of Coordinate Retrieval Systems	24
2.5	Index Vocabulary Illustration	25
2.6	Inverted File Illustration	26
2.7	Equivalent Logical Notation Illustration	30
2.8	DXT Matrix-Assignment of Terms to Documents	32
2.9	Inverted TXD Matrix, and Sample Inquiries	32
2.10	Illustration of Inclusive and Exclusive Retreival	34
3.1	Contingency Table Representation of DRS Corpus Classification by an Inquiry	45
4.1	Theoretical Versus Actual Number of Documents Retrieved -- A. D. Little Model	51
4.2	Normalized Term Usage Vs. Rank	54
4.3	Document-Term Association Matrix	54
4.4	MEZ and Zipf -- Term Frequency of Use Vs. Term Rank Distribution	61

Figure		Page
5.1	Test System Term Frequency of Use Vs. Term Rank Distribution	73
5.2	Relationship Between Term-Document Matrix and Term Usage and Cumulative Percent Utilization of Thesaurus	75
5.3	Term Frequency of Use Vs. Rank for Test Sample, Test System and Two Cases Analyzed by Litofsky (90)	77
5.4	Term Frequency of Use Vs. Rank for System Investigated by A. D. Little (1, 2)	78
5.5	Term Usage Vs. Cumulative Thesaurus Utilizations of Thesaurus for Systems Investigated by Houston and Wall (68)	79
5.6	Term Usage Vs. Cumulative Utilization of Thesaurus for Systems Investigated by Wall (143)	80
5.7	Term Usage Vs. Cumulative Utilization of Thesaurus for Systems Investigated by Litofsky (90)	82
5.8	Term Usage Vs. Cumulative Utilization of Thesaurus for Test System	83
5.9	Comparison of MEZ Canonical Form with Test System	89
5.10	Test System Depth of Indexing Density Distribution	92
5.11	Depth of Indexing Distribution for Systems Investigated by Litofsky (90)	93
5.12	Derivation of the TXT Matrix	94
5.13	Illustration of a Sparse DXT Matrix	97
5.14	Theoretical TXT (i,j) Prediction Factor-γ	103
5.15 to 5.30	Plotted γ-Factors for Test Sample for $1 < f(i), f(j) \leq 32$	104 to 119
5.31 to 5.37	Plotted Cumulative Frequency of Occurrence of the Ratio of Actual to Theoretical-γ's	121 to 127

Figure Page

5.38 Term Co-Occurrence Values for f(i) and f(j) 131

5.39 } Density of Term Co-Occurrence Values for f(i) 132

5.40 } 133

5.41 Upper and Lower Bound Limits for the γ-Factors for the Test System 138

5.42 Adjusted γ-Factors for the Test System 141

5.43 Theoretical Probability of at Least One Co-Occurrence for Terms with $f(i) = 1$ and $1 \le J_x \le D$ 150

5.44 Cumulative Plot of Quantity Output for Coefficient of Association (G) for the Test System 156

6.1 Theoretical Family of Curves Defining the Lower Bound of the Probability of Co-Occurrence of Two Terms with $f(i) = 1$, $1 \le f(j) \le D$ and $1 \le J_x \le D$, $J_y = 1$ 162

TABLE OF CONTENTS

ABSTRACT iii

ACKNOWLEDGMENTS v

LIST OF TABLES vii

LIST OF FIGURES ix

Chapter 1 INFORMATION STORAGE AND RETRIEVAL: BACKGROUND ISSUES 1

- 1.1 Introduction 1

Chapter 2 COORDINATE INDEX DOCUMENT STORAGE AND RETRIEVAL SYSTEMS: A FORMAL DESCRIPTION 13

- 2.1 Document Storage and Retrieval Systems 13
- 2.2 Document Selection: Sizing the Collection 13
- 2.3 Indexing -- Document Analysis and Representation 17
  - 2.3.1 Coordinate Indexes 19
- 2.4 The Index File 23
- 2.5 Inquiry Formulations 27
  - 2.5.1 Inquiry Grammar 28
- 2.6 Search Files and Retrieval Process 29
- 2.7 DRS -- A Brief Formal Description 36
- 2.8 Retrieval Set Characteristics 37

Chapter 3 RETRIEVAL QUANTITY AND DRS PERFORMANCE MEASURES 39

- 3.1 Introduction 39
- 3.2 Measures for Evaluation 40
  - 3.2.1 Response Time 40
  - 3.2.2 System Costs 41
  - 3.2.3 System Convenience of Use 42
  - 3.2.4 System Flexibility 43
  - 3.2.5 Retrieval Quality 44

Chapter 4 RETRIEVAL QUANTITY ESTIMATION: LITERATURE REVIEW AND PROPOSED METHODOLOGY 48

4.1 Introduction 48
4.2 General Critique of Previous Research 48
4.3 Proposed Methodology for Developing the R_q Measure 58

4.3.1 Fundamental DRS Relationships 59
4.3.2 Inquiry Definition and Generation 63
4.3.3 Inquiry -- Retrieval Quantity Measure Relationship 63

4.4 Hypotheses for Retrieval Quantity Estimations 65

Chapter 5 THE RETRIEVAL QUANTITY MEASURE: EXPERIMENTS AND RESULTS 69

5.1 Introduction 69
5.2 Setting and Description 70

5.2.1 Experiments and Analysis 72

5.3 Document Retrieval Systems -- Common Characteristics 72

5.3.1 The Term-Frequency-of-Use Distribution 74
5.3.2 The Term-Frequency-of-Use Canonical Form 81
5.3.3 Depth of Indexing Distribution 90
5.3.4 The Term-Term Co-Occurrence Distribution 91

5.4 The Retrieval Quantity Measure 120

5.4.1 Application of the Term Co-Occurrence Factor, γ and Determination of R_q 128
5.4.2 Testing the R_q Estimate 135

5.5 The Likelihood of Non-Zero Term-Term Co-Occurrences 145
5.6 Word Association Coefficients 152
5.7 System Growth Impact on Retrieval Quantity 153

Chapter 6 CONCLUSION AND SYNTHESIS OF FINDINGS 159

6.1 Introduction 159

6.2 General Conclusions 159
6.3 Management and Design Aids 161

Chapter 7 RECOMMENDATION FOR ADDITIONAL RESEARCH 169

7.1 Introduction 169
7.2 Corpus Homogeneity and Heterogeneity 169
7.3 Distribution of Terms with Common Frequencies of Use 170
7.4 The MEZ Canonical Form 170
7.5 Depth of Indexing Distribution 171
7.6 Higher Order Term Associations 172
7.7 R_q Model Extensions 173

7.7.1 Psychological Analogies 174

BIBLIOGRAPHY 176

APPENDIX A Glossary of Terms 185

APPENDIX B Institute of Library Research Laboratory DRS 188

o Thesaurus (Sample)
o Document Descriptions (Sample)

APPENDIX C Sample Data Base Characteristics 198

o Term Frequency of Use Listing
o Depth of Indexing Listing
o Term-Document Matrix in Condensed Array Format

APPENDIX D Illustrations of Computations to Estimate Retrieval Quantity 207

> It is in the nature of the mind to forget and in the nature of man to worry over his forgetfulness....
>
> Bower

Chapter 1

INFORMATION STORAGE AND RETRIEVAL: BACKGROUND ISSUES

1.1 INTRODUCTION

Man has always employed some means of storing and retrieving information. In early tribal or closed society environment man's memory was the principal repository of knowledge, the link between successive generations and between the discovery of new knowledge and those who would use it. The advent of formal speech and recordable languages provided the means for accumulation of experience and knowledge in mediums for transmission, storage and use by others, in a relatively time independent sense (93). As the scope and content of information became more voluminous and complex, formal systems were constructed for information storage and retrieval.

This report is concerned with certain underlying principles that characterize a large class of formal information storage and retrieval systems. Throughout the discussion that follows, at the risk of terminological monotony, the term information will be continuously used to describe what "it" is that information storage and retrieval systems store and retrieve. No definition of information is given, principally because there is no generally accepted precise definition available. Descriptively information has been labeled; the essential ingredient of conversation, writing and thought; recorded experience essential for decisionmaking; the essential link between means and ends; a resource; meaningful data; the result of a process on data; and a symbol or signal that a system can employ to guide or control its functions (6, 26, 27, 149). Information, however, is not considered to be knowledge, per se, or communication. On the other hand, knowledge is thought of as an organized body of information, and communication is viewed as information transfer. The notions become even more confounded when one considers the additional (though fuzzy) distinctions between data and information, data and knowledge, and so on.

Suffice it to say, that the entities -- data, information, and knowledge are different, relative in place and time, and that the basis of distinction is in part rigorously quantitative (viz., Information Theory (145)) and qualitative (i.e., contemporary, intuitive concepts and usage). For this analysis, information is intuitively treated as existing in graphic records* (e.g., documents) and to be perceivable by an inquiring mind which has a need for information.

Contemporary society can be viewed as an enormous information generating, processing, storage and retrieval mechanism. The problem of overabundance of information is compounded by a seemingly cultural magpie-like behavior which seeks to store the better part of all information and to retrieve it as well (29). There is no accurate census of the literature population, but a number of statistical estimations have been made. De Solla Price (41) has estimated that 350,000 scientific papers are published annually. Bourne (12, 13) has estimated that there are 30 to 35,000 journals published annually of which 15,000 are significant,** and that the volume of significant** papers published throughout the world per year is between 900,000 and 2,100,000. Further there are an estimated

---

*The specification of graphic records is for the purposes of this analysis, and is not meant to imply that written/printed language is the only source of information. Other media, often less restrictive, are the non-graphic verbal and non-verbal.

**No definition of significance is given.

3500 abstracting and indexing services in the world (circa 1960). In addition to the periodical population there are the monograph and abstract files. Figure 1.1 illustrates the estimation of the file size of books and periodicals in U.S. colleges and universities, and Fig. 1.2 the file size for U.S. public libraries. An estimate of the number of technical literature abstracts and/or citations produced annually throughout the world is given in Fig. 1.3.

While the per annum volume of periodicals, abstracts and monographs is impressive, the estimated growth rates are staggering. DeSolla Price (42, 43) has plotted (see Fig. 1.4) the growth of scientific and abstract journals published from the oldest surviving periodical* to the year 2000, and an exponential growth is clearly evident. Hold (67) surveyed the growth of the professional literature in economics, electrical engineering, physics, psychology and biology, and also observed exponential growth characteristics; his results are shown in Fig. 1.5. Holt (67), Brookes (19), and Krauze (79) all suggest that the growth of literature in terms of the number of articles and journals is of the form:

$$V_t = V_o e^{rt} e^{\varepsilon t} - D_t$$

where V_t = total volume of literature (in the field of interest) at time t

V_o = volume of literature at time t_o

r = the growth rate (estimated to result in a doubling every 10 years). Note: $e^{10r} = 2 \quad \therefore \quad r = 7$ percent per annum

---

* Philosophical Transactions of the Royal Society of London (1665).

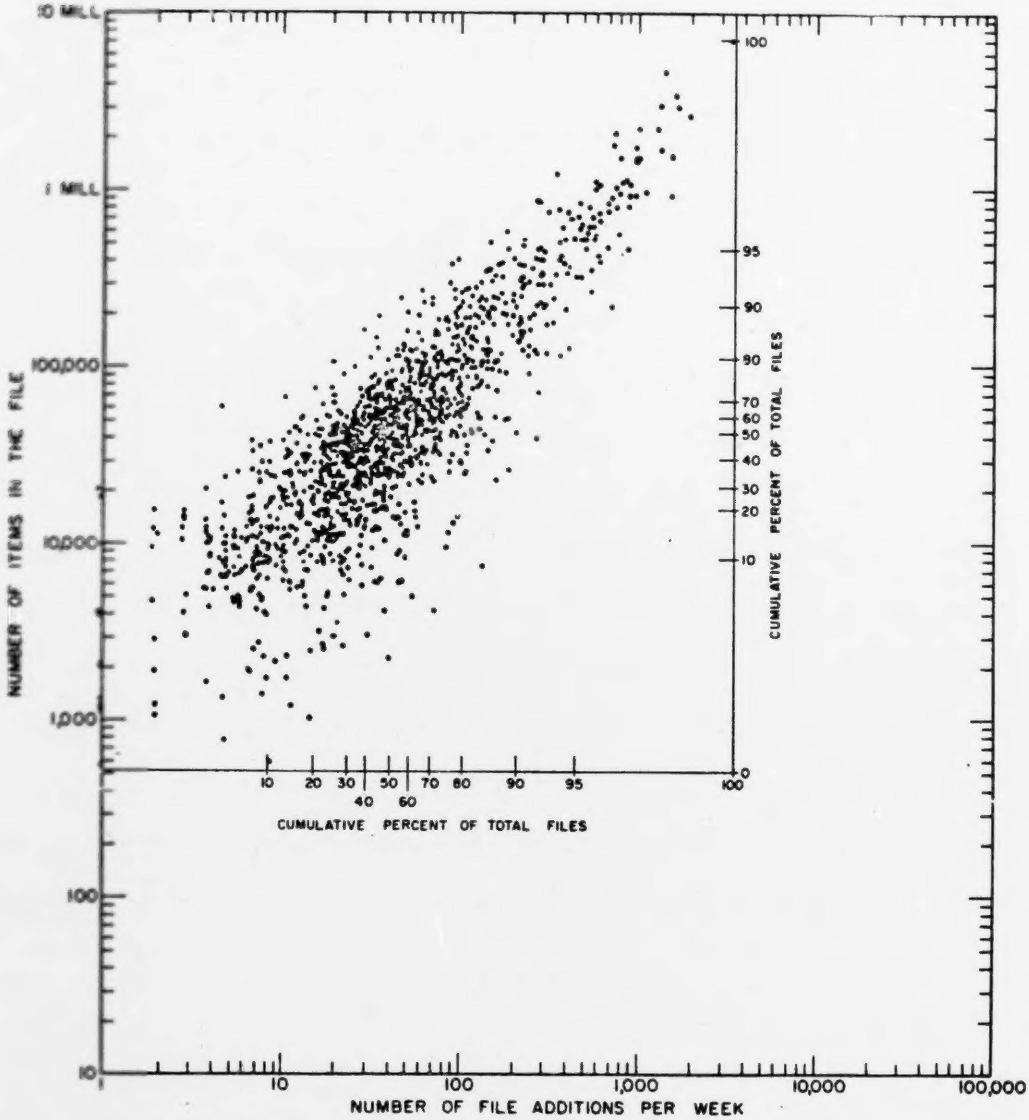


FIG. 1-1 U.S. college and university libraries—file size and accession rates. [Source: *Library Statistics of Colleges and Universities, 1959–60; Part I: Institutional Data,* U.S. Dept. of Health, Education and Welfare, Office of Education, J. C. Rather and D. C. Holladay, Report OE-15023 (1961).] (13)

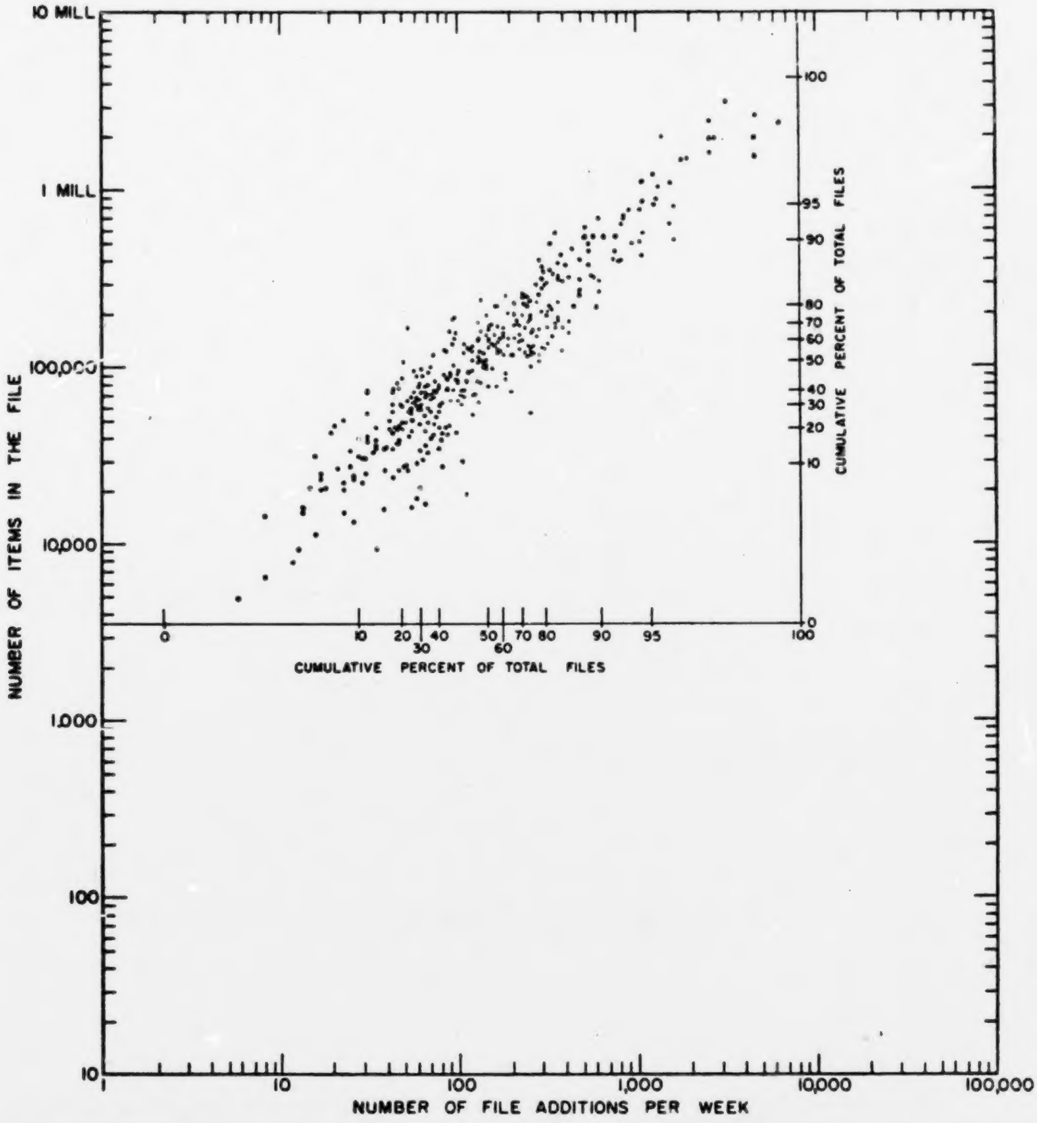


FIG. 1-2 U.S. public library systems—file size and accession rates. [Sources: *Statistics of Public Library Systems in Cities with Populations of 100,000 or More: Fiscal Year 1958*, U.S. Dept. of Health, Education and Welfare, Office of Education, Circular 590 (June 1959); *Statistics of Public Library Systems in Cities with Populations of 50,000 to 99,000: Fiscal Year 1958*, U.S. Dept. of Health, Education and Welfare, Office of Education, Circular 594 (July 1959); *Public Library Statistics: 1944–45* (for cities with populations of 25,000 to 49,999), Federal Security Agency, Office of Education (1947).] (13).

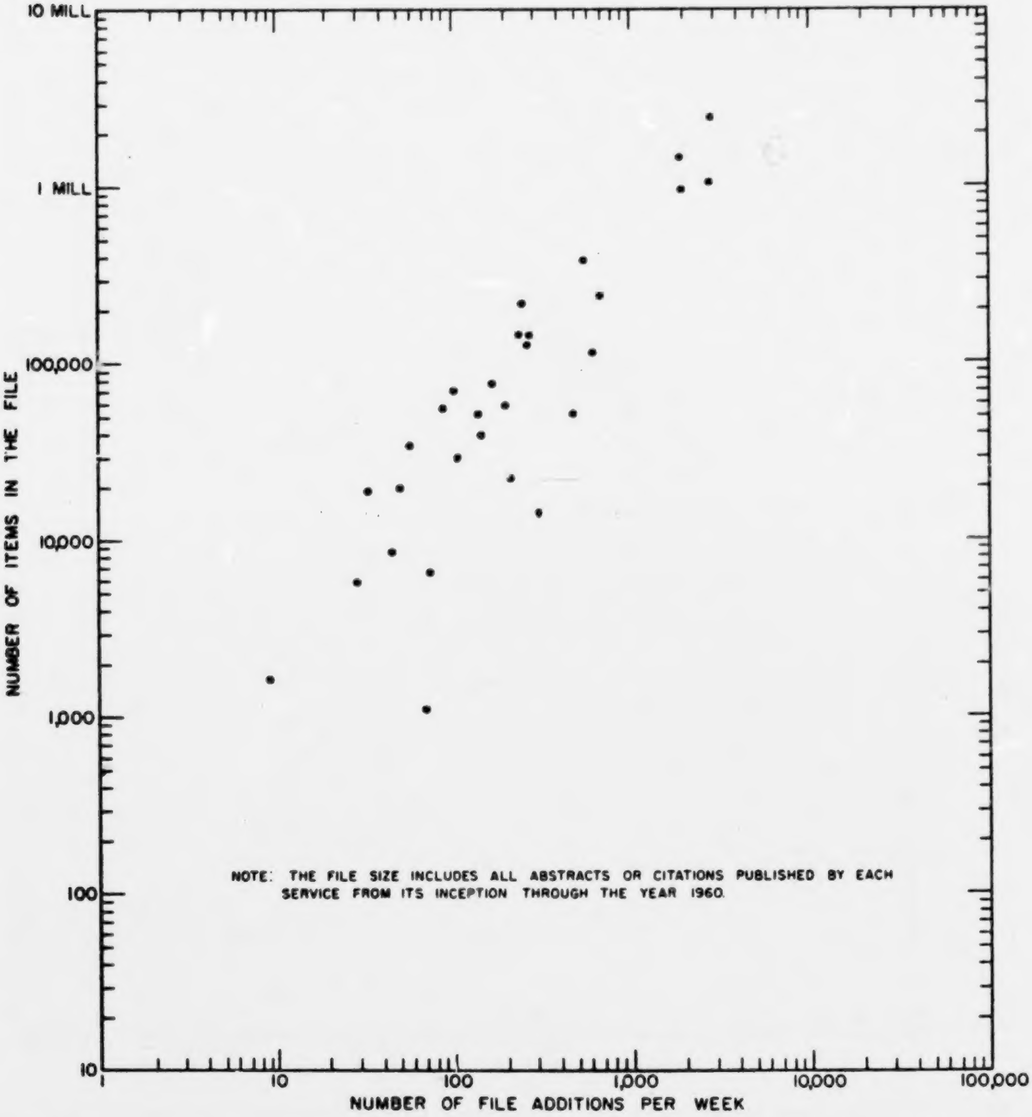


FIG. 1-3 Accumulated file sizes and current accession rates of the publications of several abstracting and indexing services. *Note:* The file size includes all abstracts or citations published by each service from its inception through the year 1960. (13).

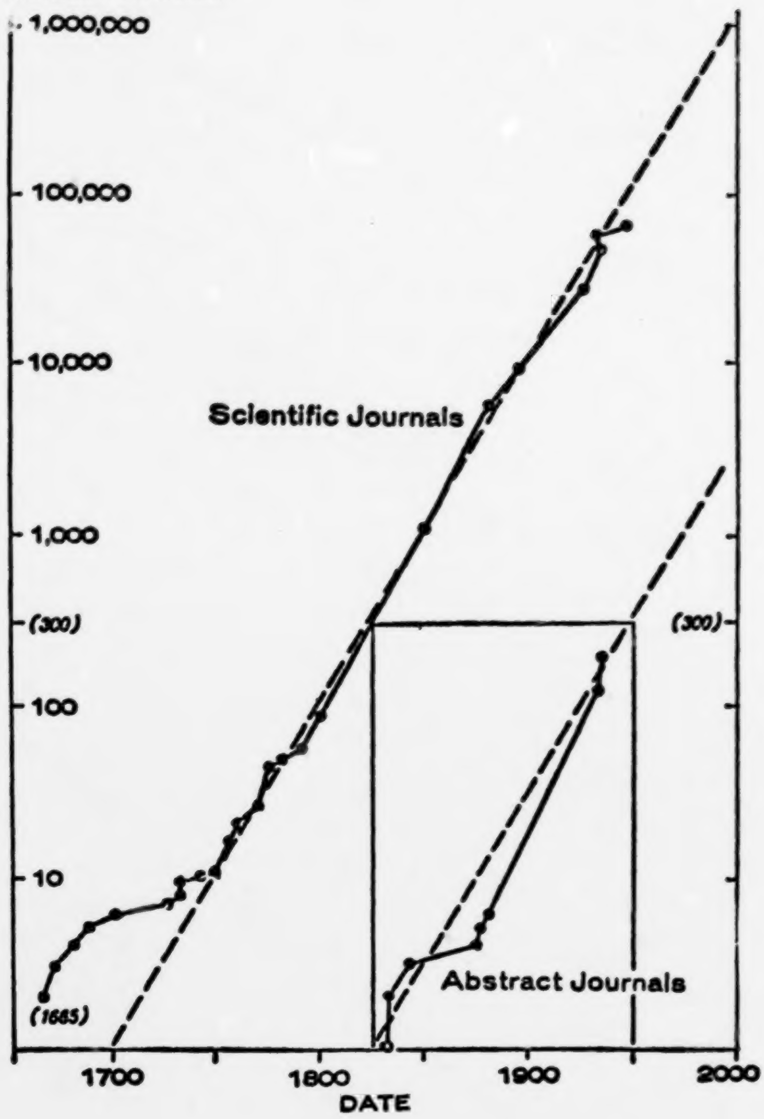


Figure 1.4

Growth of Journals and Abstract Journals (43).

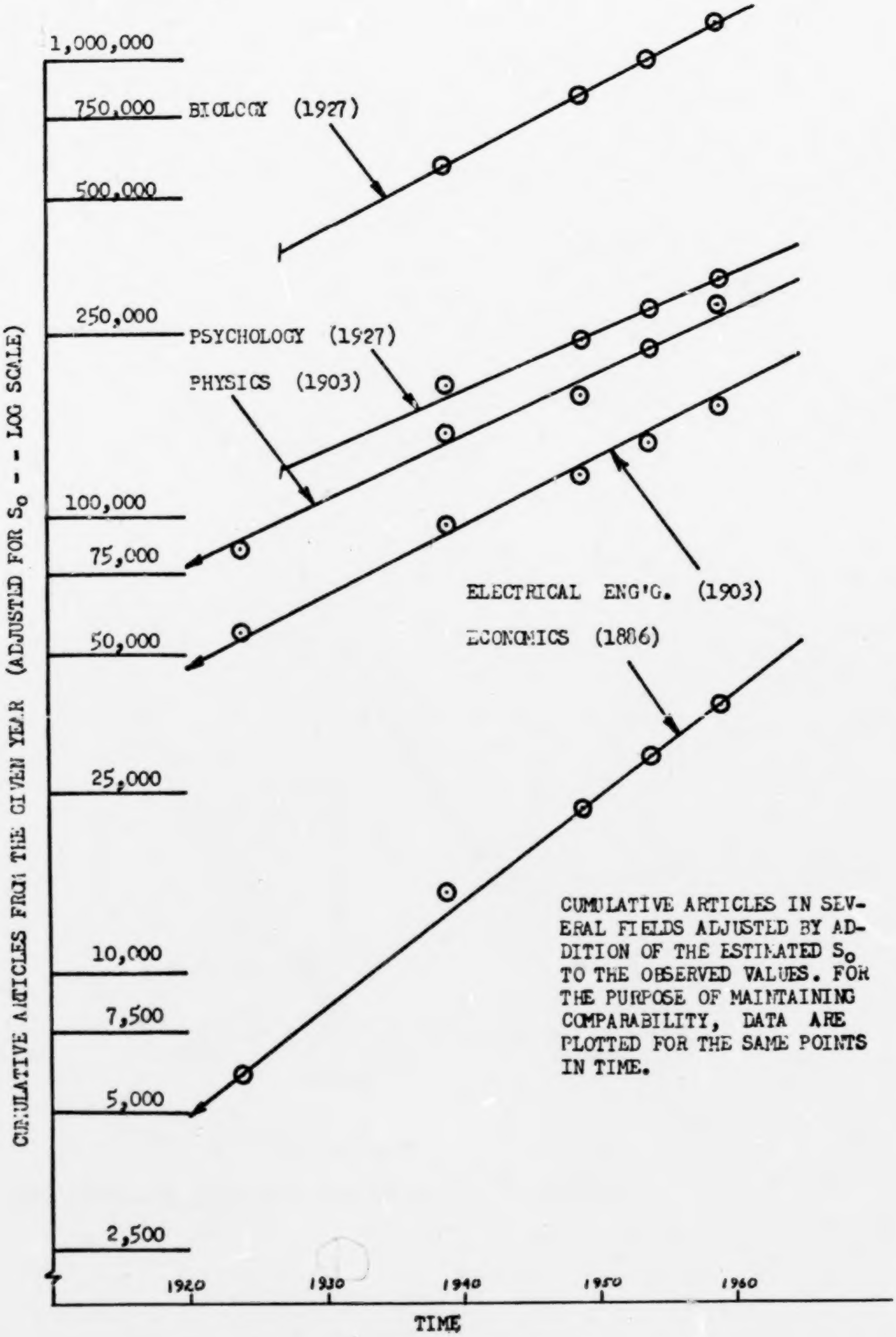


Fig. 1.5 -- Periodical Growth (67)

ε_t = the statistical error of measurement; assumed to have the property Expected Value $(\varepsilon_t) = 0$

D_t = the death rate of journals, and the death rate or near-usefulness of articles.*

The above analysis, admittedly cursory and non-rigorous, does imply an information control, storage and retrieval problem. All evidence seems to say that for any established field there is an abundance of information, and it is growing.

There is a bonefide need to store a substantial portion of existing literature, and there is a need for a physically feasible means of retrieving information that is both economically practical, and time and content relevant to the information user. Concern about this information handling problem has placed new emphasis on the traditional activities of assemblying and coding recorded information, and has resulted in the emergence of a new discipline, Information Science, which focuses on the analysis and solution of information, storage and retrieval (ISR) problems. A variety of systems, processes and techniques has been constructed to cope with many ISR problems, and a typical set of ISR processes and their interactions are illustrated in Fig. 1.6.

An important subset of ISR systems are document retrieval systems (DRS), which, as the title implies, retrieve documents and hence the information in them indirectly. This subset of systems, for instance, excludes fact retrieval or intelligence retrieval systems. The term

---

*A great deal of conjecture surrounds the assessment of D_t. It is believed (Brooks (19)) that it has exponential properties, but these are very relative to the user and subject in question.

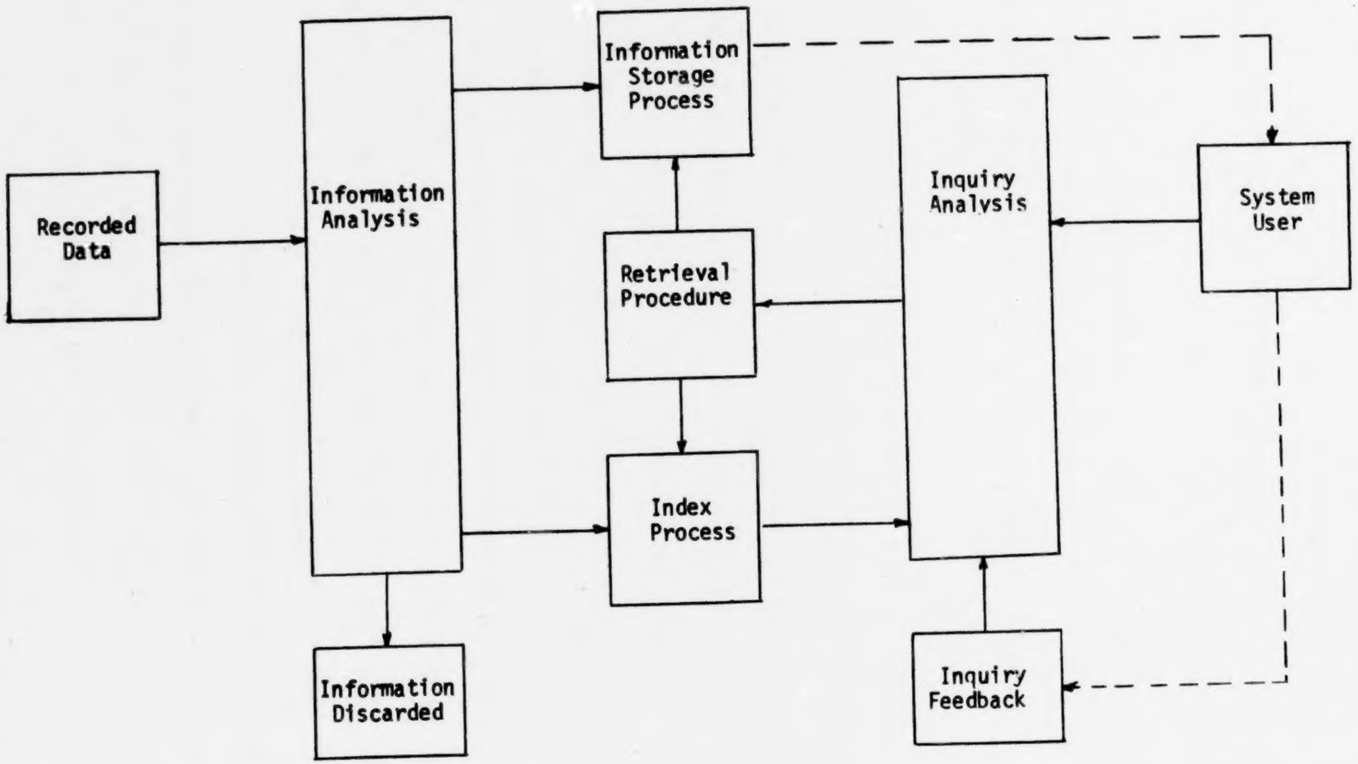

Recorded Data
Information Analysis
Information Discarded
Information Storage Process
Retrieval Procedure
Index Process
Inquiry Analysis
Inquiry Feedback
System User

Fig. 1.6 -- Information storage and retrieval functions

document is used as a generic for information bearing items -- monographs/books, periodical articles, abstracts, film, machine coded/readable tape, etc. It is with this particular class of ISR systems that this report is concerned,

To date, extensive research has been carried out on various aspects of DRSs. Ostensibly, major efforts* have been made in index analyses and evaluation by Cleverdon (30, 31, 32), Taube (135, 136), Gull (61), Thorne (138) and Swanson (128); user satisfaction by Borko (11), Bourne (14, 15), Fairthorne (47), Goffman (56), Rees (112, 113) and Swets (131, 132); Retrieval Output relevancy by Barhydt (5), Cuadra (36, 37, 38), Doyle (45), Goffman (57), Lancaster (82), and Salton (117, 118, 119); and, automatic classification by Litofsky (90); notwithstanding these and other efforts, more problems remain unsolved than solved in the design, management and evaluation of DRSs. Of particular interest is the class of problems concerning the estimation of the retrieval quantity of DRSs. This particular dimension of DRS performance has not been thoroughly analyzed, and no satisfactory operational solution has been suggested.

The basic objective of this research is to develop a methodology that will enable designers and managers of DRSs to estimate the quantity output in response to an inquiry, prior to the processing of the inquiry. A secondary objective is to demonstrate how the estimation methodology can be used to assess DRS changes over time.

---

*No attempt is made to be exhaustive, the cited work is intended to be a representative sample of previous efforts by some of the more well-known researchers in Information Science.

Before proceeding with the derivation of the retrieval quantity (R_q) measure, the context and qualifications of the analysis will be presented. In the next chapter the specific class of DRSs for which the R_q estimation procedure is to apply are described, and Chapter 3 presents a survey of the many DRS measures of performance to place the R_q measure in perspective.

Chapter 4 presents a discussion of previous efforts to develop an output quantity measure, and also contains a formal description of the recommended methodology to develop the R_q estimate. In Chapter 5, a description of the experiments performed to evaluate the R_q estimation procedure, and the results of the experiments are presented. Chapter 6 presents a discussion of various applications of the R_q measure to aid in the management and design of DRSs. Also, Appendix A contains a glossary of terms to Information Storage and Retrieval terminology.

> A man should keep his little brain attic stocked with all the furniture that he is likely to use, and the rest he can put away in the lumberroom of his library, where he can get it if he wants it.
>
> Sherlock Homes

Chapter 2

COORDINATE INDEX DOCUMENT STORAGE AND RETRIEVAL SYSTEMS: A FORMAL DESCRIPTION

2.1 DOCUMENT STORAGE AND RETRIEVAL SYSTEMS

Document Retrieval Systems (DRSs) are a class of information retrieval systems solely concerned with the subject analysis of document content, the storage of a set of official surrogates "defining" document content, and the "mechanical" search of the surrogate set to identify or select those documents most "relevant" to a user's formal request. The basic functions of a DRS are illustrated in Fig. 2.1.

Of special interest to this discussion are system output, user inquiries, and index characteristics. Since each of these processes and products is embedded in a system and is directly influenced by other system components, a brief review of the major system functions will be presented to place following developments in proper system perspective.

2.2 DOCUMENT SELECTION: SIZING THE COLLECTION

Mention has already been made of the existing volume and growth of documented information, and of the associated problems of researchers, students, etc. concerned with keeping abreast of their fields of interest.

It is elementary, however, to note that not all existing information related to any one subject should be stored in DRSs serving

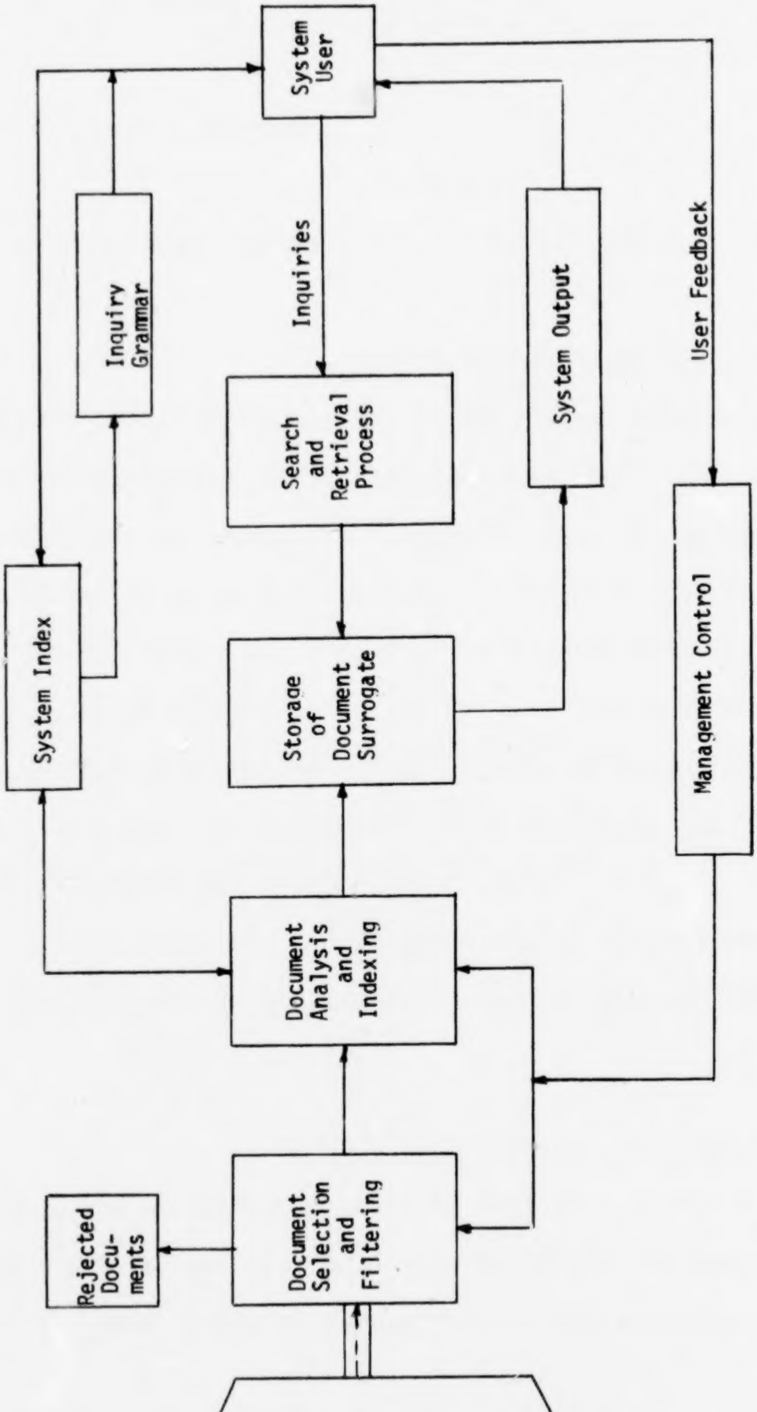

Documented Information
Document Selection and Filtering
Rejected Documents
Document Analysis and Indexing
System Index
Storage of Document Surrogate
Inquiry Grammar
Search and Retrieval Process
Inquiries
System User
System Output
User Feedback
Management Control

Fig. 2.1 -- Document storage and retrieval system functions

users in that field. As well it is equally evident, that not all newly generated documented literature is a contribution to that field, and uncurbed storage of documents would result in unsatisfactory DRS performance. From the point of view of the user the quantity and quality of systems output would leave much to be desired. From the point of view of the manager, the costs of indexing, analysis and searching would be out of balance with the systems effectiveness. In order to manage a document collection, selection criteria are required and document filtering are necessary. Simply put, not all documents in a subject field should be input into a DRS, and not all documents input in the DRS should be stored forever.

With regard to the issue of document collection size there are certain models that have been developed that can aid the DRS designer and manager to estimate the number of documents or journals that should be reviewed to yield a desired number of subject-relevant "documents," or conversely to estimate the number of "documents" that are generated by a certain number of journals. The two models have been referred to by Leimkuhler (88), as the Bradford Law of Scattering and the Bradford Law of Distribution, and as one might suspect are inversely related. Bradford (16) first stated the relationship of "documents" to journals. as follows:

> If a large collection of papers is ranked in order of decreasing productivity of papers relevant to a given topic, three zones can be marked off such that each zone produces one-third of the total of relevant papers. The first, the (sic) nuclear zone, contains a smaller number of highly productive journals, say n_1; the second zone contains a larger number of moderately productive journals, say n_2, and the outer zone a still larger number of journals of low productivity, say n_3. The Law of Scatter states that,

$$n_1:n_2:n_3 = 1:a:a^2$$

where a is a constant.

In the subject of geophysics, which Bradford analyzed, "a" was approximately equal to five.

Subsequent to Bradford's effort Vickery (141), Kendall (75), Leimkuhler (88), Fairthorne (49) and Brookes (20) have each made contributions to the interpretation and operationality of the Bradford Law of Scatter. Leimkuhler (88) has shown the inverse relationship between the Law of Scatter (the distribution of the number of journals containing a given fraction of relevant documents) and the Law of Distribution (the distribution of document productivity in a collection of journals) and has expressed the latter in the following form:

$$F(x) = \frac{\ln(1+\beta x)}{\ln(1+\beta)}$$

where $F(x)$ = the cumulative fraction of "documents" in a collection of journals on a specific subject

x = the corresponding fraction of the most productive journals in the collection; and $0 \leq x \leq 1$.

β = a constant related to the subject field and the completeness of the journal collection.

The above model enables a DRS designer or manager to estimate the relationship between the number of documents in the system corpus, and the number of documents in the population of journals on a specific subject. In other words, the Bradford relationships can be used* to relate the productivity of a collection of journals to the population of journals, and aid in the selection of journals to yield documents for the corpus. Given a subject field and budget constraints, these relationships can aid in the cost/benefit tradeoff between budget dollars and the number of documents/journals to collect.

2.3 INDEXING -- DOCUMENT ANALYSIS AND REPRESENTATION

For this discussion, indexing will be defined as the assignment of subject content indicating terms to a document. The purpose of the indexing operation is to make it possible to search a file of the content indicating terms, that are mapped onto the set of documents, as a substitute for searching the document set, and to identify those documents relevant to an inquiry. Relevant is used here to mean that condition in which the terms used in the inquiry are also used to describe the selected documents.

It is of course theoretically possible to review the set of documents as opposed to the index file, but this approach quickly becomes physically and economically impractical for even moderate collections (several hundred) of documents. Thus the index provides a manageable set of content indicating terms and classes to be searched in place of the corpus, and provides a vehicle to identify those documents in the corpus most likely to contain the desired information.

There is in fact a spectrum of indexing philosophies, and associated techniques with various proper names. That they are all related or relatable has been discussed by Artandi (3), Bourne (13), Jahoda (71), and demonstrated by Foskett (52). Basic to any indexing process is the set of vocabulary terms employed to describe the content of the documents. The set of vocabulary terms constitutes the index language, and as well, an important part of the inquiry language of DRSs. The latter property follows from the fact that once the index terms have been assigned to the set of documents, they are then used to represent the documents and become the vehicle to map inquiries onto the corpus.

---

*Groos (60) has observed a departure from the linear relationship in log-log space of the Bradford Law when plotting the Keenan-Atherton data for physics. The observed deviation, however, has not been thoroughly evaluated to determine if the cause lay in the assumptions of the Bradford Law or in the incompleteness of the experimental observations.

Traditionally, subject classification concepts involve the use of formal schemes to organize the subject matter in a predetermined order to some prescribed depth of detail. Typically, these traditional classifications are hierarchical in nature; that is, there exists among the set of descriptors a rather precisely defined relationship of every term to every other term. At the other end of the spectrum there are the "key word" systems, which in their simplest form have no word relationships defined, and usage of -- and addition to -- the descriptor vocabulary is unrestricted. Artandi (3) makes a useful distinction between "systems vocabulary" and "lead-in-vocabulary" as a means of distinguishing between word indexing and subject indexing. They are both methods of representing document content, but they differ operationally. By "systems-vocabulary" it is meant the set of terms under which document content descriptor entries are made; that is, the terms used to index the documents. The "lead-in-vocabulary" of a DRS, "is an index referring from terms used in the literature to terms in the system vocabulary, (3)." The principle characteristic of word indexing is that descriptors or words are employed as they are found in the text of documents to serve as index terms. Thus word indices are derived from the documents that are being indexed.

The key word in context (KWIC) index is an example of word indexing, in its simplest form, involving elementary alphabetical permutations of the "key words" in the document titles.

2.3.1 Coordinate Indexes

Word indices in which the index terms are manipulated or coordinated are called coordinate[*] index systems. Further, those DRSs in which the coordination of the descriptors is done in the indexing process are called pre-coordinate DRSs. Analogously, those systems in which the coordination of the descriptors takes place during the inquiry generation process are called post-coordinate DRSs. The pre- and post-distinction obviously refer to the temporal occurrences of the event of combining descriptor terms.

The important characteristic of pre-coordinate DRSs is that the searching occurs, and the inquiries generated, using the terms and their combinations the indexor has prepared. There is no additional coordination of the descriptors at the time of the inquiry.

Traditional examples of pre-coordinate systems are the hierarchical systems in which a tree structure is employed to define a generic-subordinate relationship and the coordinated relationships among the subordinate terms. Figure 2.2 illustrates a typical hierarchical scheme, and some examples are the Library of Congress, Dewey Decimal, and Universal Decimal Classification Systems.

---

[*] As first developed, "coordinated terms" literally implied the statistical conjunction of two or more terms. However, the meaning of "coordinate index" as used in most post-coordinate-index systems has been broadened to incorporate the full set of Boolean operators, and in some instances even syntactical, semantic and syndetic term-relationships.

Another class of pre-coordinated systems are facet indices. A facet is a set of terms which occurs with sufficient frequency in a subject field to provide a useful category or facet of terms for the description of documents in that field. A schematic of a facet index is given in Fig. 2.3. In these systems, the pre-coordination of the descriptor terms occurs at the time the facet is defined. The concept of faceted systems for subject description was first developed by Ranganathan (110) in his colon classification scheme.

Although the above two classes of pre-coordinate index systems exhibit strong structural properties, there are also pre-coordinate systems which have no hierarchies or proper set structure. Such systems essentially consist of a set of descriptors (the vocabulary), and a set of indexing and vocabulary control rules.

Post-coordinate index schemes, as noted previously, are exemplified by the combination of more or less elemental index terms at the time of inquiry generation and search initiation. These systems are adaptive in that they can accommodate shallow or deep indexing as well as simple or complex inquiries. In their earliest form, post-coordinate retrieval systems were known as Uniterm systems, after Taube (135). The uniterm is a unit or elemental concept, usually a single word, used to describe the subject of a document. In many systems, the vocabulary is quite often derived from the text and title of the documents to be indexed, and no control is applied over the vocabulary or the coordination of the descriptor terms. The post-coordinate index system is a

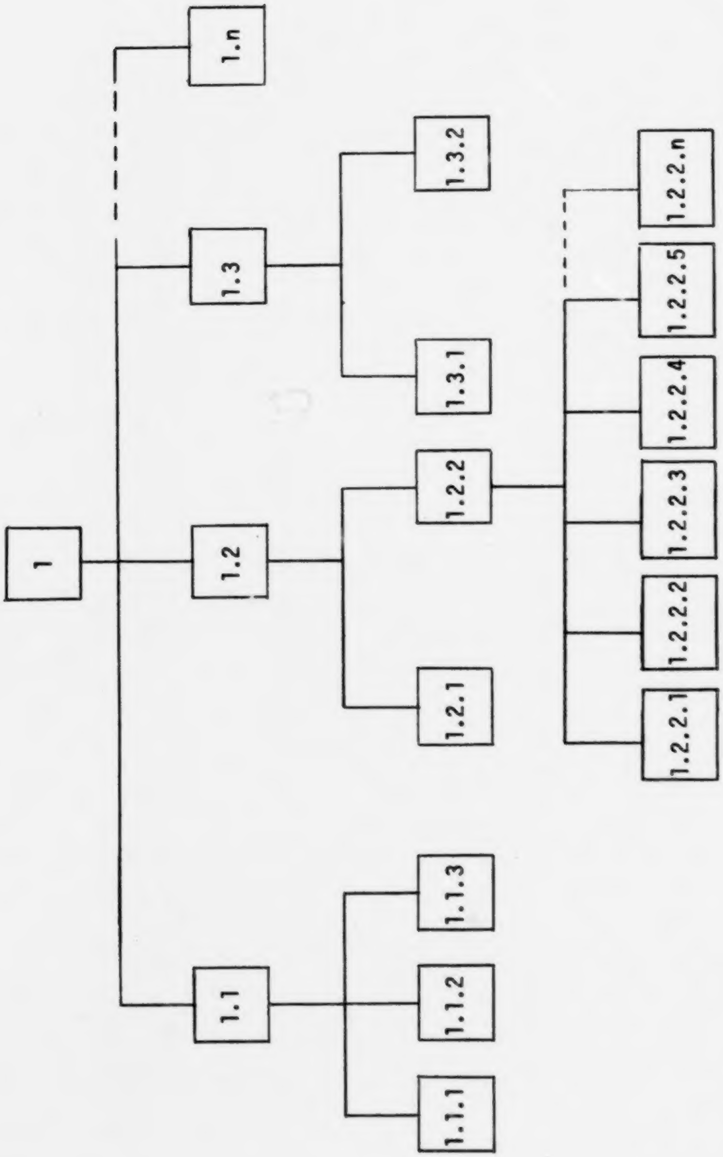

1
1.1
1.2
1.3
1.n
1.1.1
1.1.2
1.1.3
1.2.1
1.2.2
1.3.1
1.3.2
1.2.2.1
1.2.2.2
1.2.2.3
1.2.2.4
1.2.2.5
1.2.2.n

Fig. 2.2 -- Illustration of a hierarchical, tree of knowledge, structure

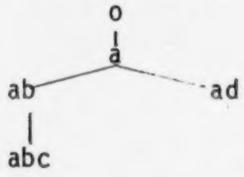
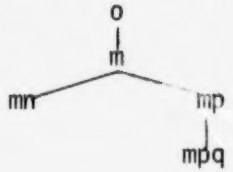
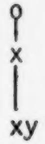
Facet	Element (Terms)	Predecessor	Tree Facet Diagram
	a	o	
1.	b	a	
	c	b	
	d	a	
	m	o	
2.	n	m	
	p	m	
	q	p	
	x	o	
	y	x	

Illustration of Facets

It should be noted, that any element in Facet 1 could be followed by any element in Facet 2 or 3; and, any element in Facet 2 could be followed by any element in Facet 3.

Fig. 2.3

very versatile scheme and can be adapted to incorporate a broad set of characteristics. Figure 2.4 illustrates a taxomony of coordinate retrieval systems, and various logical extensions to other types of index systems. Of central relevance to this report are the post-coordinate Document Retrieval Systems that incorporate Boolean operators in the system language.

2.4 THE INDEX FILE

The index file in a coordinate index system consists of the descriptor/index vocabulary and the descriptor tracings or assignments to the documents in the corpus. A sample of an actual index vocabulary for the subject area of Information Science, is given in Fig. 2.5, and a sample of a term frequency of use ranking is presented in Fig. 2.6.

Of particular interest are the following characteristics of a coordinate index system file:

(1) the number of active terms in the vocabulary

(2) the frequency of use of each term

(3) the depth of indexing for the documents in the corpus

These characteristics are indicative of the term-document distribution in the DRS which is the basic relationship in these systems. It is important to realize that all these characteristics are dynamic in nature. They will change as new index terms are added, or created out of combinations of existing terms, and as new documents are added to -- and old documents dropped from -- the corpus. The index vocabulary is used by the system user to generate, in a post-coordinate sense, inquiries to the DRS.

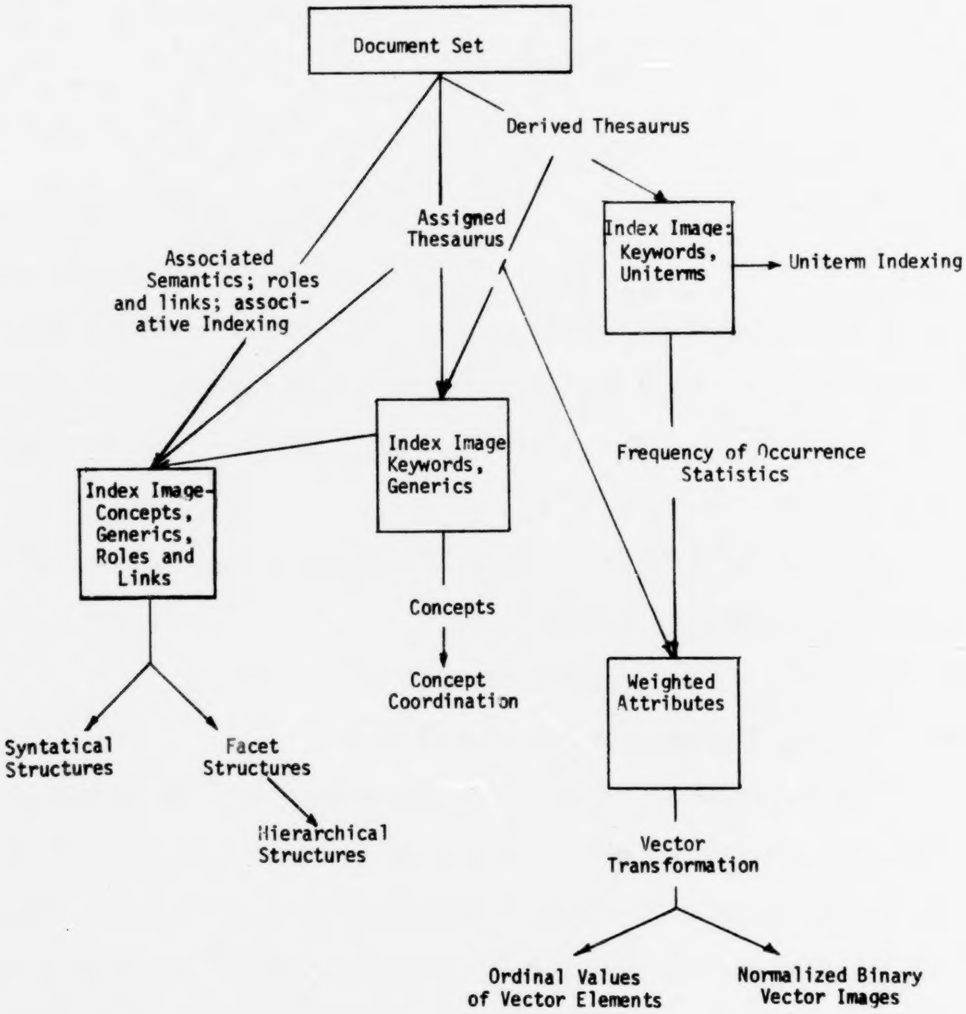

Document Set
Derived Thesaurus
Assigned Thesaurus
Associated Semantics; roles and links; associative Indexing
Index Image: Keywords, Uniterms
Uniterm Indexing
Index Image Keywords, Generics
Index Image-Concepts, Generics, Roles and Links
Frequency of Occurrence Statistics
Concepts
Concept Coordination
Weighted Attributes
Syntatical Structures
Facet Structures
Hierarchical Structures
Vector Transformation
Ordinal Values of Vector Elements
Normalized Binary Vector Images

Fig. 2.4 -- Coordinate index models

```
ABBREVIATIONS

     S  = SEE
     SA = SEE ALSO
     SN = IN THE SENSE OF (I.E. SCOPE NOTE)
     *  = NO DOCUMENTS YET INDEXED WITH THIS TERM
     +  = TERM NOT ALLOWED, RELATED TERM TO BE USED
```

```
*ABBREVIATION
 ABSTRACT
 ABSTRACTING
 ACCESS
 ACCESSION NUMBER
 ACCURACY
 ACQUISITION
 ADDRESS
 ADMINISTRATION
 ALGEBRA
+ALGOL
       S    PROG. LANGUAGE
 ALGORITHM
 ALPHABETIC
 ALPHABETIC ORDER
 ALPHANUMERIC
*ALTERNATIVES
 AMBIGUITY
 ANALOGY
 ANALYSIS
 ANSWER
*ANTHOLOGY
       SA   BIBLIOGRAPHY
 APPLICATION
+ARITHMETIC
       S    MATHEMATICS
 ARRAY
+ARTICLE
       S    DOCUMENT
 ARTIFICIAL INTEL
 ASSIGNED
 ASSOCIATION
 ASSOCIATIVE

+ATTRIBUTE
       S    CHARACTERISTIC
 AUTHOR
 AUTHORITY LIST
       SA   THESAURUS
 AUTO ABSTRACTING
 AUTO. INDEXING
 AUTOMATIC
 AUTOMATION
       SA   MECHANIZATION

 BATCH PROCESSING
 BIBLIOGRAPHIC
 BIBLIOGRAPHY
       SA   ANTHOLOGY
 BINARY
 BOOK
 BOOLEAN
       SA   LOGICAL

 CALL NUMBER
 CANONICAL
       SA   NORMALIZED
 CARD
 CARD CATALOG
 CATALOG
 CATALOGING
 CATEGORIES
 CENTERS
 CENTRALIZED
 CHARACTERISTIC
```

Fig. 2.5 -- Index Vocabulary Illustrations (from Maron (98))

INDEX TERM	NO. OF REFS.
INFO. RETRIEVAL	84
SYSTEM	84
DOCUMENT	78
COMPUTER	69
STORAGE	69
INDEXING	64
RETRIEVAL	63
INFORMATION	59
SEARCHING	58
ANALYSIS	53
CLASSIFICATION	52
STRUCTURE	52
INDEX	49
RELEVANCE	49
LANGUAGE	46
EVALUATION	44
EXPERIMENT	44
ASSOCIATION	42
SEMANTIC	41
MATRIX	39
NATURAL LANGUAGE	38
WORD	36
FREQUENCY	35
DESCRIPTOR	34
QUESTION	33
DICTIONARY	32
PROGRAM	32
USER	32
DATA	31
MEASURE	31
TRANSLATION	31
LIBRARY	30
RELATIONSHIP	30
THESAURUS	30
HIERARCHY	29
ALGORITHM	28
AUTOMATIC	28
COMMUNICATION	28
INPUT	28
LINGUISTIC	28
STATISTICAL	28
SYNTAX	28
PROBABILITY	27
GRAMMAR	26
OUTPUT	26
QUESTION-ANSWER	26
REFERENCE	26
WORD ASSOCIATION	25
LITERATURE	24
FILE	22
LOGIC	22
MATCH	22
PROCESSING	22
RELEVANT	22
SEARCH STRATEGY	22
SYMBOL	22
TECHNICAL	22
AUTO. INDEXING	21
BIBLIOGRAPHIC	21
SCIENTIFIC	21
STAT. METHOD	21
CONCEPT	20
EFFICIENCY	20
RECALL	20
TEXT	20
THEORY	20
ABSTRACT	19
CO-OCCURRENCE	19
CODING	19
KEYWORD	19
TRANSFORMATION	19
WEIGHT	19
GRAPH	18
VOCABULARY	18
CLUMP	17
HARDWARE	17
MODEL	17
SUBJECT	17
SYNONYM	17
SYNTACTIC ANAL.	17
TREE	17
COMPARISON	16
COORDINATE INDEX	16
CORRELATION	16
MECHANIZATION	16
TAG	16
TEST	16
ACCESS	15
BIBLIOGRAPHY	15
CLASSIF. SCHEME	15
CONTENT	15
COST	15
EDUCATION	15
LATTICE	15
LINK	15
MATHEMATICAL	15
RETRIEVAL SYSTEM	15
TITLE	15
ASSOCIATIVE	14
MEANING	14
NETWORK	14
RESEARCH	14
SCANNING	14
SERVICE	14
ABSTRACTING	13
BOOLEAN	13
CITATION INDEX	13

Fig. 2.6 -- Index term list sorted on frequency of use (from Maron (98))

2.5 INQUIRY FORMULATIONS

The fundamental components of inquiry formulations are--the user's need for information, the systems inquiry vocabulary (the index file), and the system inquiry grammar.

The notion of user need for information is principally psychological in nature; it is very dynamic and directly dependent on the relative state of knowledge of the user. The reason for noting the user's need at this point is primarily to identify the source of the DRS workload or demand. The expressing of a need for information, in the terms and grammatical structure of the system, is the system inquiry. It is usually the case that the formal inquiry is only a partially accurate representation of the "real" need on the part of the user. However, for the purposes of this analysis the formal inquiry will be taken as the complete system workload, as the system output variable of interest is quantity. The knotty issues of distinguishing between felt-need, expressed request and formal inquiry and their respective "noise" contribution to the relevance* and nonrelevance of systems output are not dealt with.

The fundamental components of the formal inquiry are the descriptor terms incorporated in the inquiry, and the grammatical operators used to "coordinate" the terms. The descriptor terms have been described, and the grammar used in DRSs will be discussed next.

---

*There has been more analysis related to the concept of relevance--its definitions, measurement and quantification than any other Information Retrieval System characteristic. To mention just a few, see Cooper (33,), Barhydt (5), Cuadra and Katter (36, 37, 38), Doyle (45), Salton (112), Swets (131), Swanson (129), and Maron and Kuhns (97).

2.5.1 Inquiry Grammar

The operational manner in which the descriptor terms are coordinated in an inquiry is defined by the system grammar. Of the class of Information Storage and Retireval systems that this analysis deals with, the nature of the grammar is quite primitive; only certain explicit operations/connections are permitted, between system controlled vocabulary terms, in the "coordination" process.

The formal representation of coordinate retrieval system grammars can take several forms. A common representation is in terms of a logical language, for example, a sentential or propositional calculus. In this analysis, the rules of term combinations can be formally represented by a Lattice Algebra,* or its less general proper subset, Boolean Algebra.** In the rest of the discussion a Boolean Algebra structure will be assumed. Essentially, the specifications of the relationship between two classes of objects is what Boolean Algebra is all about. Very briefly, this structure, for a defined set T and its elements (A, B,..), is defined in terms of the following operations.

Conjunction; $C = A \cdot B$, the subset or subclass of all index terms or elements of T that are both in the subsets of A and B.

---

*Excellent presentations of Lattice theory are provided in Birkhoff (9) and Szasz (134), Applications to DRS theory can be found in Becker and Hayes (6) and Salton (117).

**A Boolean Algebra is defined as a distributive lattice in which each element "a" has a complement defined by its negation.

Disjunction; D = A + B, the subset of all index terms or elements of T which are either in subset A or subset B.

Negation;* N = -B or $\bar{B}$, the subset of all index terms in T which are not in subset B.

Figure 2.7 illustrates many of the different symbolic and graphical notations in use to represent the above logical operations. For this analysis the notations "." for conjunction, "+" for disjunction and "-" for negation will be employed consistently.

In sum, the inquiry language (grammar and vocabulary) is the vehicle to translate user's information needs into formal system inquiries. Subsequent to the generation of the request, the next step is the search and retrieval proccss.

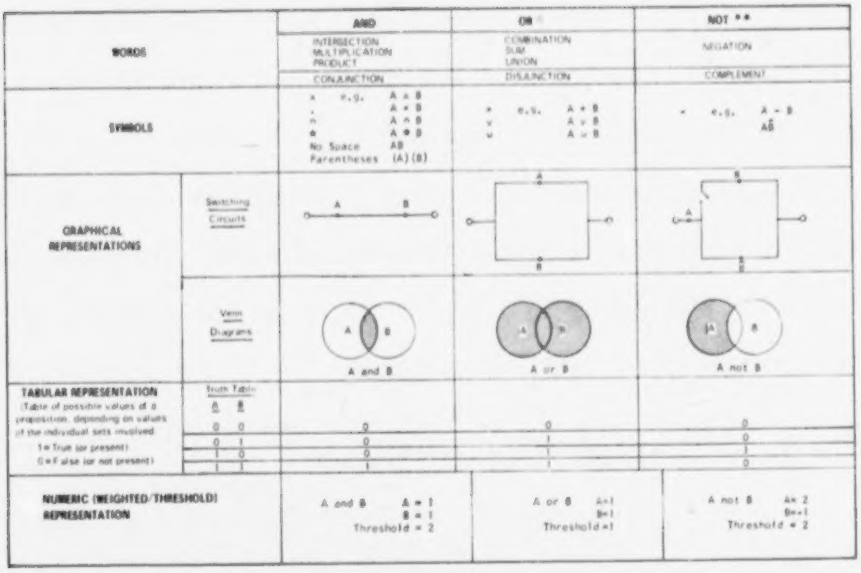
2.6 Search Files and Retrieval Process

A central DRS component is the storage or search file,** which contains the descriptions of corpus documents. This file provides the means whereby formal requests are compared with the index descriptions of the documents. In a sense, there is an input indexing operation (on the documents), and an output indexing operation (on the user's request). Given that both requests and documents are represented by

---

*There are variations on the operation Negation that can be used in DRSs; for example, Praternegation--implicit exclusion instead of explicit exclusion, Soergel (125); Brouwerian Compliment--the smallest set of items that with certainty contains all the NEGATED elements, Salton (117); Psuedo Compliment--the largest set of items that with certainty contains no NEGATED elements, Salton (117).

**In actuality, most DRSs have two search files; one for the document descriptor images, and one for the physical storage of the documents. Only the former files are of concern here.



		AND	OR *	NOT **
WORDS		INTERSECTION MULTIPLICATION PRODUCT	COMBINATION SUM UNION	NEGATION
		CONJUNCTION	DISJUNCTION	COMPLEMENT
SYMBOLS		× e.g. A × B · A · B ∩ A ∩ B * A * B No Space AB Parentheses (A)(B)	+ e.g. A + B ∨ A ∨ B ∪ A ∪ B	− e.g. A − B A$\bar{B}$
GRAPHICAL REPRESENTATIONS	Switching Circuits			
	Venn Diagrams	A and B	A or B	A not B
TABULAR REPRESENTATION (Table of possible values of a proposition, depending on values of the individual sets involved 1 = True (or present) 0 = False (or not present)	Truth Table A B			
	0 0	0	0	0
	0 1	0	1	0
	1 0	0	1	1
	1 1	1	1	0
NUMERIC (WEIGHTED/THRESHOLD) REPRESENTATION		A and B A = 1 B = 1 Threshold = 2	A or B A=1 B=1 Threshold=1	A not B A= 2 B=-1 Threshold = 2

* "OR" IS COMMONLY UNDERSTOOD AS THE "INCLUSIVE OR", I.E. A OR B OR BOTH. ALSO UTILIZED IS THE "EXCLUSIVE OR," I.E. A OR B BUT NOT BOTH, SOMETIMES SYMBOLIZED AS ⊕

** "NOT" IS EQUIVALENT TO "AND NOT" NOT "OR NOT"; IT IS SOMETIMES EXPRESSED "BUT NOT".

Fig. 2.7 -- Equivalent logical operations and notation (from Brandhorst (17))

lists of index terms, the retrieval process consists of matching the two lists, and retrieving those documents whose descriptions sufficiently overlap or match the inquiry.

The assignment of index terms to documents can be represented in matrix form. A hypothetic assignment of terms to documents is shown in Figure 2.8. In this example, the index terms are represented by the set T, and the set of documents by D, where T and D also represent the power of the respective finite sets and are usually not equal. As indicated in the example, the form of the term to document assignment is a binary operation, represented by the blank or zero and 1 notation; the latter representing assignment. While other assignment operations are possible, notably weighted assignments, the more common index operations are binary, and will be the type assumed in this analysis.

The search file can be represented in matrix (DXT) form, with the columns constituting the index term profile of the corpus, and the rows representing the membership of documents to the index term or concept sets. There are two basic arrangements for the search file, index term on documents (TXD) or the inverted file shown in Figure 2.9, and documents on terms (DXT) as shown in Figure 2.8. The DXT arrangement is the usual output from the indexing operation, and the TXD (the transpose of DXT)is the more convenient form for searching and retrieving documents. The retrieval process consists of a subject search of the document descriptions. Several simple cases of subject searches are illustrated in Figure 2.9. Search request (1) is a simple one descriptor inquiry, which would retrieve three documents. For this kind of search request those documents that belong to index sets defined

Set of Index Terms
(T)

Set of Documents
(D)

	T_1	T_2	T_3
D_1	1	1	0
D_2	0	1	0
D_3	1	1	1
D_4	1	0	0
D_5	0	1	1
D_6	1	0	1

Fig. 2.8 -- DXT matrix -- assignment of terms to documents

	D_1	D_2	D_3	D_4	D_5	D_6
T_1	1	0	1	1	0	1
T_2	1	1	1	0	1	0
T_3	0	0	1	0	1	1

Inquiry	Output
I. T_3	D_3, D_5, D_6
II. T_2 and T_3	D_3, D_5
III. T_1 and (T_2 or T_3)	D_1, D_3, D_6

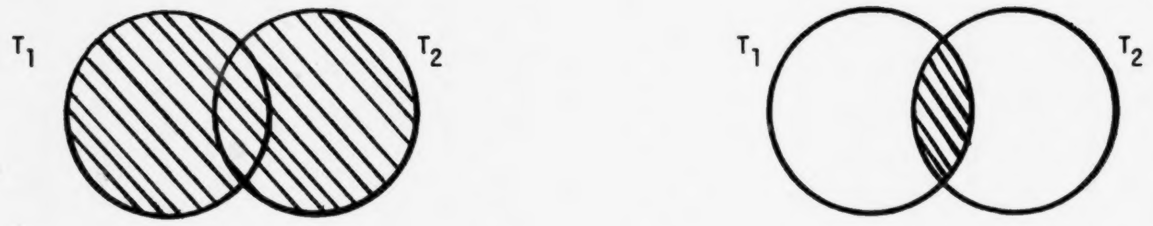
Fig. 2.9 -- Inverted TXD matrix, and sample inquiries

in the inquiry are retrieved, regardless if other index descriptors are also assigned to the specific documents.

There are different retrieval strategies that can be used in coordinate index DRSs to select inquiry "relevant" documents. The two major strategies to be considered are direct match and word associations retrieval. The simplest direct match request is the single term inquiry, already noted. The next and more common request is the conjunctive coordination of two or more terms. These logical product inquiries require that the documents retrieved have all the inquiry terms assigned as subject descriptors, and the search result is defined as an exclussive mapping on the search file. That is, only those documents dealing with the inquiry "exclusively" are retrieved. Figure 2.10 illustrates an exclusive search by logical statements and Venn diagrams.

A less restrictive direct match request is to disjunctively coordinate descriptors as a logical sum. In this type of inquiry each term is treated as a logical equivalent or synonym of every other term, and any document description containing one or more terms is retrieved. These logical sum inquiries result in an inclusive mapping on the search file. For the same set of inquiry terms, the inclusive search output will contain the exclusive search set. An illustration of an inclusive search logic is given in Figure 2.10. In general, inquiries will contain combinations of logical products and sums of index terms, and occasionally, negation of a term. Term Negation is treated in this analysis as the compliment of the logical product operation.

The second retrieval strategy is word association searching, in which the initial inquiry is expanded or broadened so as to retrieve more documents in the corpus that are "relevant" to the initial inquiry.



T_1 or T_2

$T_1 + T_2$

Inclusive Inquiry

T_1 and T_2

$T_1 \cdot T_2$

Exclusive Inquiry

Fig. 2.10 -- Illustration of inclusive and exclusive retrieval

Association retrieval techniques are based on the relationships between descriptor terms assigned to the DRS corpus. There are basically four categories of word relationships that can be used as a basis for inquiry term augmentation: (1) Semantic relationships which manifest the meaning and context of terms within a language, (2) Syntatic relationships which arise from terms as members of word classes and with the class relationships in a structural (grammatical) sense, (3) Syndetic relationships which measure the manner by which words that are conjunctively coordinated with a given or base term cross-reference one another, and (4) statistical relationships which measure the frequency of occurrence of terms in a document.

For this analysis, only the statistical association will be discussed in that it is the most common technique for inquiry modification. The emphasis (in later chapters) will be on their operational definition. As implied by the name, statistical term association does not address the semantic, syntatic or syndetic connections among terms; rather, it views terms as separate isolatable units and is based principally on the frequency of terms usage within a given DRS corpus. The basic assumption is that, within the context of a given corpus, terms which are statistically correlated with one another are presumed to be meaningfully associated. Hence the implication is that if terms A and B were determined to be associated, for a given corpus, and term A appears in a inquiry that inquiry could be expanded by the disjunctive incorporation of term B to term A. The objective of including term B is to increase the likelihood of retrieving a larger set of inquiry "relevant" documents from the corpus.

2.7 DRS -- A BRIEF FORMAL DESCRIPTION

The above discussions have been basically informal, and it is instructive to consider what a formal statement of a DRS consists of. The advantages of a formal statement are: (1) that the elemental or basic components of the system and their relationships are defined, so as to provide a sound basis for intra-system analysis, and (2) to facilitate inter-system structural and operational comparisons.

Formally,* a coordinate index DRS is defined as consisting of:

1. A set of distinct documents to be analyzed/indexed

$$D = \{d_1, \ldots d_D\}$$

2. A set of elementary descriptors/attributes/index terms from which compound-descriptors (combinations) can be constructed

$T = \{t_1, \ldots, t_T\}$ -- the elemental set of attributes

$T' = \{t'_1, \ldots, t'_T\}$ -- the set of terms generic to set T and composed of combination of elements in T

$\vdots$

3. A set of statements/axioms which connect descriptors with documents. This set of statements defines a homomorphic mapping between the set of descriptors T and the set of documents D. The mapping usually results in a binary set of assignments,

---

*For extensive definitions of formal systems see Curry, et al. (39), and for an excellent discussion of a formal system definition of DRSs see Soergel (125).

$$\{T\} \Rightarrow \{D\} : \text{DXT (binary)}$$

but it is not necessarily restricted to 0 or 1 assignments; weighted assignments are also possible.

4. A set of statements (theorems) derived from the axioms and the system grammar which define the manner of coordination and relationship of descriptors for searching and inquiry specifications.

2.8 RETRIEVAL SET CHARACTERISTICS

It follows from the preceding discussions that the properties of the retrieval set are a function of three parameters:

(1) the number of terms and the degree and type of coordination in the inquiry

(2) the search strategy -- either direct match or word association

(3) the DRS DXT distribution -- from which all the DRS characteristics can be derived.

The retrieval set characteristics are definable in terms of quantity and quality. The quality measure is a reflection of the user's judgment of the relevance of the retrieved material. The quantity measure is simply the number of documents output in response to the inquiry, and is the retrieval set characteristic of interest to this discussion.

The principle task is to define the quantity output as a function of the above noted parameters; inquiry, search strategy and the DXT distribution. Various hypotheses about the functional relationship and the parameters will be presented and analyzed in Chapters 4 and 5. However, before addressing those issues, a statement of how retrieval quantity is related to existing DRS performance measures is necessary to provide additional perspective for the measure as a management and design tool.

Performance measures like sign posts guide the way....

Chapter 3

RETRIEVAL QUANTITY AND DRS PERFORMANCE MEASURES

3.1 INTRODUCTION

In this section, the need for a Retrieval Quantity (R_q) measure will be discussed, and the relationship of the proposed measure with other DRS performance measures will be noted.

The tasks of design and management of DRSs require tools and performance measures to aid in the selection of preferred candidate options, and in the control over the fundamental processes of inquiry analysis, indexing, retrieval and system output. The designer needs tools that reflect the cause-effect relationships between the DRS building blocks of thesaurus, corpus and term-document distribution. Before a DRS is built, the design should be assessed and compared to alternative designs. Existing DRSs require management tools to tune the system to meet the needs of the user, and to control the changes in the system due to growth in the thesaurus and corpus. Users of DRSs need guidelines to construct and adjust inquiries to more completely meet their information needs, both in quantity and quality.

Some of the tools and performance measures are available, and a basis for an overall analytic framework also exists, although a rigorous systems formulation has yet to be developed. A brief survey of a number of the measures that can be used for design and management will be presented next, and the R_q relationship to the different measures briefly noted.

3.2 MEASURES FOR EVALUATION

The primary purpose of a DRS is to cost-effectively over time provide the system users with the information requested when it is needed. The major dimensions of evaluation implied in this objective statement include: time, cost, flexibility, convenience of use, information-quality, and information quantity.

3.2.1 Response Time

In general, for information systems, the dimension of time reflects the period to perform an operation such as providing the user with a response to an inquiry.* Lowe (92) and Hayes (63) have investigated various time processing distinctions between different file organizations for storage and retrieval operations. Also, it follows that the amount of time to process an inquiry will be proportional to the thesaurus size and term frequency of use distribution. In fact, Webster (145) has demonstrated that certain DRS dictionary searching techniques are critically affected by the term frequency of use distribution. In many DRSs the requests are batch processed, and from the user's point of view the response time is fixed. However, the amount of "processing" time is still of interest to the system manager. In those systems in which there is an on-line real-time environment, the user, by necessity, also becomes acutely aware of processing times.

One possible way to anticipate required inquiry processing time is to use the inquiry as a basis for estimating the required search and retrieval operations. The procedure for predicting the retrieval quantity measure (to be presented in Chapter 4) entails a set of iterative steps proportionate to the "complexity" of the inquiry. Assuming a balanced file and dictionary look-up scheme in which each step takes approximately the same amount of time to process, by estimating the retrieval quantity, and keeping track of the number of iterations required, the user and manager could gauge the inquiry processing time and workload demands, respectively.

---

*A more restrictive definition of response time is offered by Lancaster and Climenson (84) who define it as the average time required to obtain a satisfactory response from the system.

3.2.2 System Costs

Various recommendations have been made for measures of cost-effectiveness for information storage and retrieval systems. Overmeyer (105) has published a relatively detailed cost analysis of the American Society of Metals System of Western Reserve University. Lancaster (85) discusses relevant system factors susceptible to cost-analysis, and suggests possible tradeoffs between input and output costs and between alternative candidate DRSs. Tell (137), Kochen (78), Bryant (23), Westat (147) and Lancaster (84), have developed DRS cost-analysis models of various degrees of detail. Notwithstanding these efforts, a comprehensive operational model for costing still remains to be developed. A sound basis for DRS cost analysis appears to exist; for example, Lancaster (85) provides a subject relevant framework that could be coupled with the concept of opportunity costs and a well-developed system analysis setting, as in Fisher (51). It appears, however, that standard cost accounting methods cannot be conveniently or correctly carried over to DRS operations. As Marron (99) notes, a corpus of documents is not really like or analogous to equipment or machinery, particularly with regard to the concept of depreciation or amortization. Also, the costs and effort of constructing a corpus are not very sensitive to the demand volume for services. As well, the problem of correctly tracing input and operational costs is particularly difficult when there are several information services performed by the system; for example, dissemination, retrieval, abstracting, etc. Also, most DRSs operate in a non-market setting in which the users of the system do not "pay" for the service, and the system does not "compete" to provide the service. This situation tends to complicate the costing of resources consumed and the estimation of benefits accrued.

To some degree, the retrieval quantity estimate can aid in the costing of inquiries by using the inquiry processing time estimate, noted above is multiplied by a cost per unit processing time. Also, the cost estimation per inquiry can help the user "balance" his needs with the probable system accrued costs.

3.2.3 System Convenience of Use

The principal issue in the dimension of convenience of use is the amount of effort that is required from the system user to interact with the DRS. To some degree the literature on man-machine interaction has some bearing. Certainly the notion of unburdening is relevant. Investigations by Saracevic (120), Lancaster (82, 83), and Lesk and Salton (86) indicate that there is a need for user -- search analyst interaction, but there is no concensus as to whether the interaction should take place before the search or after the retrieval. There is no convenience-of-use measure of what is efficient user-system interaction.

Clearly, a fundamental parameter is the state of the user's need for information. Martyn and Vickery (100) discuss a number of conditions affecting user need, and Voigt (142) has prepared an early (1959) but still accurate description of user needs for information. It would seem that given a communicable information need, a retrieval quantity estimation process can aid in tuning the user's inquiry to the expected/desired size of the response. This notion is discussed further in Chapter 6.

3.2.4 System Flexibility

Flexibility is meant to be a measure of the DRS's capacity for positive adaptation. An implemented DRS can only stay successfully operational if it is adaptive. Of interest for this measure is what do systems have to be adaptive for, and in what ways can this flexibility be built into the system structure. Ironically, most DRSs are justified on the basis of the rapid growth and rate of change of relevant literature, and yet the systems are designed for the point in time when they are implemented, with little regard given to the need for flexibility to accommodate system growth. In addition, to the inherent growth of the corpus and thesaurus, DRSs should also have a certain flexibility to adapt to changing user needs and behavior. One of the greatest faults of the traditional library classification schemes is the implicit assumption that all library users are counterpart mini-models of the classification scheme, and as well will never change. A more preferred state is one in which a DRS would interact with users at different levels of user proficiency, and grow in a controlled sense with the incorporation of new material.

An important impact of growth is that as the corpus and thesaurus changes the system output will be different at different points in time for the same inquiry. The retrieval quantity measure can be used to gauge the impact of corpus and thesaurus growth on the DRS output quantity, and in this dimension provides a measure of system adaptability. This application of the retrieval quantity estimate is discussed in Chapter 6.

3.2.5 Retrieval Quality

Measures of retrieval quality have by far received the most attention of the DRS dimensions of evaluation. By retrieval quality it is meant the relevance, pertinence or correctness of the retrieval document information to the user's information need.

For any document corpus, only a fraction of the collection will contain relevant information regarding a specific user inquiry. For example, if there are D documents in the corpus, then only R may be relevant to the particular inquiry. Without the entire set D being retrieved, it is unlikely that all R relevant documents will be retrieved in any one search, initiated by the inquiry. Usually, only a fraction H of the R relevant documents are retrieved, and by definition $M = R-H$ will be missed. Also it is usually the case that a number of I irrelevant documents will be retrieved by the system in response to the inquiry. Following Vickery (140) these characteristics of a DRS and the retrieval set are represented in a two-by-two contingency table as shown in Fig. 3.1. For this binary construction, all the D documents in the system are accounted for, with respect to the inquiry which generated the retrieval set. Namely,

	Relevant	Not Relevant	
Retrieved	a (Good Hits)	b (Bad Hits)	a + b = H
Not Retrieved	c (Bad Misses)	d (Good Misses)	c + d = M
	a + c = R	b + d = I	D

Fig. 3.1 -- Two x two contingency table of an inquiry response (140)

a documents are good hits because, $a \subset R$ and $a \subset H$

c documents are bad misses because, $c \subset R$ and $c \not\subset H$

Presuming, of course, in this simple system that it is desirable to retrieve all R relevant documents. Also included in the retrieved set H are:

b documents which are bad hits because, $b \subset I$ and $b \subset H$

and the remaining,

d documents are good misses because, $d \subset I$ and $d \not\subset H$.

From the two-by-two contingency table in Fig. 3.1, a plethora of retrieval efficiency measures, primarily directed at assessing relevance/quality, have been derived. Table 3.1 lists a sample of the derivable measures. Fundamental to all of these measures are two variables -- a relevance judgment and the quantity of documents (relevant and/or irrelevant) output. The close relationship between output information quality and quantity in these measures is clearly evident. A predominant characteristic of these measures is that they are all designed to be computed after the retrieval operation, and consequently are of limited use to predict output or the effect of a system change. The retrieval quantity estimate is a step in the direction of developing management tools for predicting retrieval output and impacts due to system change.

A review of previous attempts to construct a Retrieval Quantity estimate, and the suggested methodology to predict R_q, developed in this analysis, are presented in the next chapter.

Table 3.1

RETRIEVAL SET MEASURES

Measure	Equation (based on Figure 3.1)
Resolution factor (106)	$\frac{a+b}{D}$
Elimination factor (106)	$\frac{c+d}{D}$
Pertinency factor (Relevance measure) (106)	$\frac{a}{H}$
Noise factor (106)	$\frac{b}{R}$
Recall factor (106)	$\frac{a}{R}$
Omission factor (106)	$\frac{C}{R}$ (Type I error)
Generality ratio (31, 32) Concentration ratio (47)	$\frac{a+c}{D}$
Fall out (69)	$\frac{b}{I}$
Specificity (113)	$\frac{d}{I}$
Distillation factor (47)	$\frac{ad-bc}{(a+b)(c+d)}$
Discrimination factor (47)	$\frac{ad-bc}{(a+c)(b+d)}$
False acceptance (101)	$\frac{b}{R}$ (Type II error)

Chapter 4

RETRIEVAL QUANTITY ESTIMATION: LITERATURE REVIEW AND PROPOSED METHODOLOGY

4.1 INTRODUCTION

The main body of this chapter is concerned with a review of past work related to retrieval quantity estimation. The second part of this chapter describes the proposed methodology for prediction of output quantity.

4.2 GENERAL CRITIQUE OF PREVIOUS RESEARCH

Surprisingly there have not been many analyses of the output quantity of DRSs; the review that follows is quite exhaustive. Though various approaches have been employed, all the research to date on the determination of retrieval quantity has, either implicitly or explicitly, been based on the assumption that index terms are used as though they are independent of one another. The general lack of qualification or modification of this assumption has been the rather pervasive Achilles' heel of the efforts to date. This is so because index terms do not occur or co-occur as though they are independent of one another. To assume that they do exhibit independence causes large divergences between actual and "theoretical" values of term co-occurrence and output quantity.

The earliest attempt to estimate retrieval quantity appears to have been by Bernier (7), in which the following argument is made. For a system of D documents, T descriptors, a uniform depth of indexing of t descriptors per document, with no two documents possessing an identical set of descriptors and indexing being an "essentially-random" assignment process, a n-term conjunctively coordinated inquiry has the following probability of retrieving at least one document:

$$P(R_q \geq 1) = D(\frac{t}{T})^n$$

This model is quite "hypothetic" due to the very restrictive assumptions which limits its usefulness. First, terms are not assigned as though they are balls being selected randomly from an urn; secondly, the depth of indexing distribution of DRSs is anything but uniform; and thirdly, for systems of even moderate size the above probabilities are so small as to provide almost no insight into the retrieval process.

A more ambitious attempt was made by A. D. Little (1, 2) in which a model to predict the average number of documents to be retrieved for a given inquiry is constructed. The expected number of documents retrieved is defined as a function of:

(1) the number of terms coordinated in the inquiry (only conjunctive inquiries were used) -- n
(2) the number of documents in the corpus -- D
(3) the average depth of indexing -- Ω
(4) the frequency of use distribution of index terms -- which is approximated by a geometric series and incorporated in the function by a factor $(1-\beta)/2, \beta < 1$
(5) the term usage distribution for users generating inquiries --S
(6) the index term correlation for indexing documents (the assumption of independent term usage with a correction factor was employed) --S

(7) the effect of a system "requestor" to aid in the specification of search inquiries (an implicit factor)

with the resulting function:

$$R_q = \frac{D}{S}\left[S\Omega\left(\frac{1-\beta}{2}\right)\right]^n \qquad \text{for } n > 2.$$

This model, though containing many system and inquiry characteristics, does not perform very well at all as shown in Fig. 4.1. The assumption of term-term independence is the principal factor. Also the assumption for the inquiry terms selection distribution, while not an essential ingredient for determining retrieval quantity, is not necessarily the same distribution as for terms used to describe documents.

A more abstract approach is suggested by Switzer (133) who employed a term-term distance measure to estimate the elements of the term correlation matrix (TXT). Switzer does not estimate the expected number of documents to be retrieved for an inquiry, but does note that once the term-term couplets are estimated, the logical extension to evaluating term combinations in inquiries is possible. The principle assumptions in this analysis are:

(1) the normalized co-occurrences are considered to be probabilities (a frequency interpretation of probability is implied)

(2) the term co-occurrences are hypergeometrically distributed

The proposed relationship for the value of the couplet of terms a and b is:

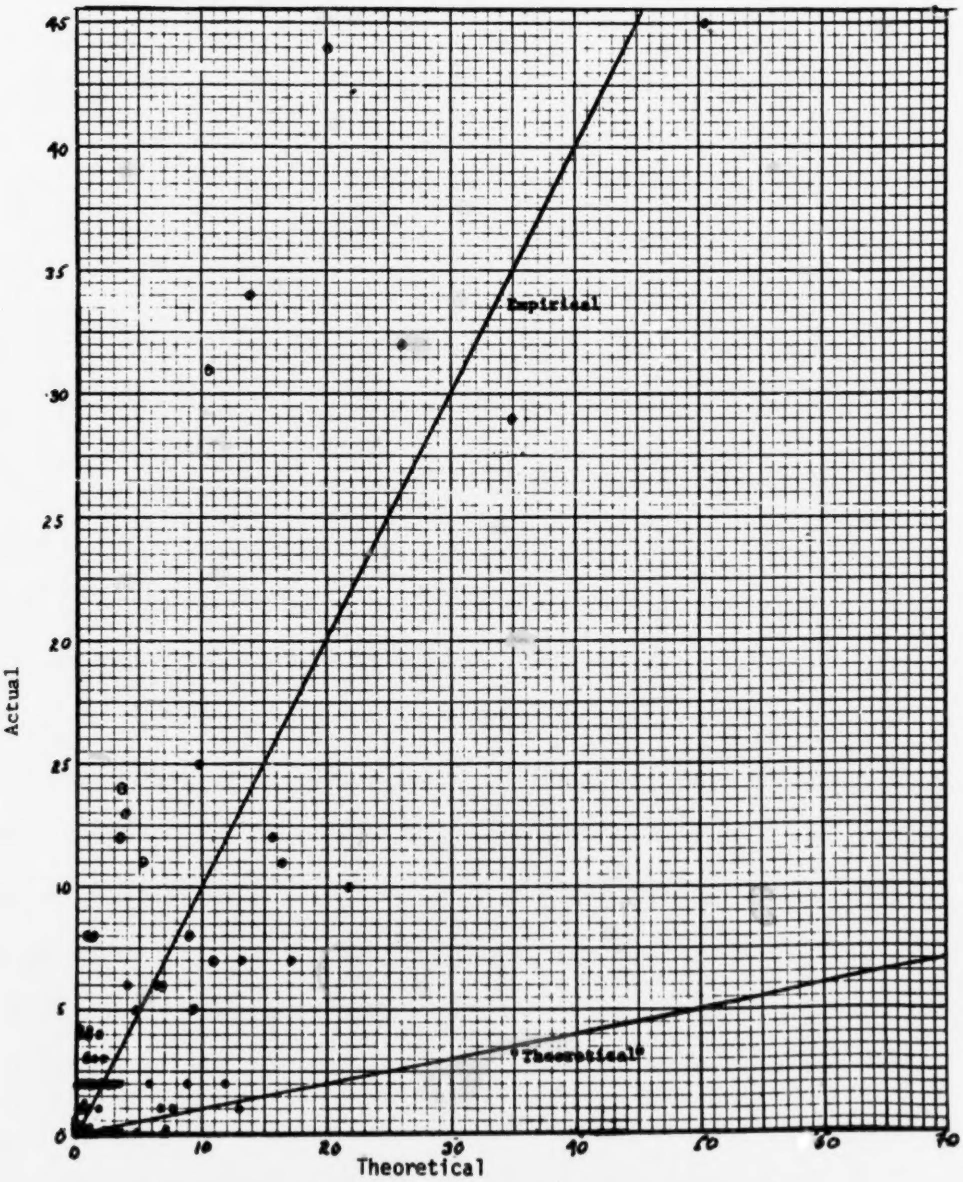


Fig. 4.1 -- Comparison of actual number of documents retrieved with theoretical number based on assumption of term-term independency - (from Ref. 2)

$$\frac{N_{ab}}{D} = \Delta_{ab} = \sum_{x=N_{ab}}^{\min(N_a,N_b)} \frac{\binom{N_a}{x}\binom{D-N_a}{N_b-x}}{\binom{D}{N_b}}$$

which is the hypergeometric distribution with parameters

N_{ab} = the number of co-occurrences for terms a and b

D = the number of documents in the corpus

N_a = the number of times term a has been used

N_b = the number of times term b has been used

Switzer did not empirically test this relationship, but it is clear from the fundamental assumptions of hypergeometricity, which is random sampling from a finite population without replacement, that it is not correct. As noted previously, term-term co-occurrences do not occur as though they are the result of a random sample.

One of the more interesting formulations to estimate document output is presented by Raver (111), in which the term frequency of use distribution is approximated by a normalized log function. The explicit distinction between a normalized and unnormalized term frequency of use distribution is very useful. In addition, Raver notes that all Boolean combinations of terms are reducible/definable by the "and" and "or" operators with those terms.

The logrithmic relationship between the frequency of term use and the term rank (in which the term with greatest use is given rank 1, the next most used term rank 2, and so on) is of the form:

$$T_r = N'\frac{T-r}{T}$$

for a normalized distribution,

where N' = most frequently used descriptor (normalized)

T = total number of active descriptor terms out of a thesaurus of size $\bar{T}$

r = rank of the term; $0 \le r \le T$ and is defined by

$$= \begin{cases} 0 \text{ when } T_r = \begin{cases} N \text{ for unnormalized distributions} \\ N/f_{min} \text{ for normalized; } N' = N/f_{min} \end{cases} \\ T \text{ when } T_r = \begin{cases} 1 \text{ for normalized distributions} \\ f_{min} \text{ for unnormalized distributions} \end{cases} \end{cases}$$

where f_{min} is the frequency of use of the least used term in the active subset of the thesaurus.

Obviously, in those systems in which $f_{min} = 1$, the term frequency of use distribution is automatically normalized. An illustration of the normalized term frequency of use distribution is given in Fig. 4.2.

From the above relationship, Raver then shows that

(a) the average number of documents per descriptor is:

$$J_o = \frac{N}{\ln_e N}$$

(b) the average number of descriptors per document is:

$$T_o = \frac{(f_{min})^N}{T}$$

(c) the average number of documents to be retrieved for an n term (conjunctive) inquiry is:

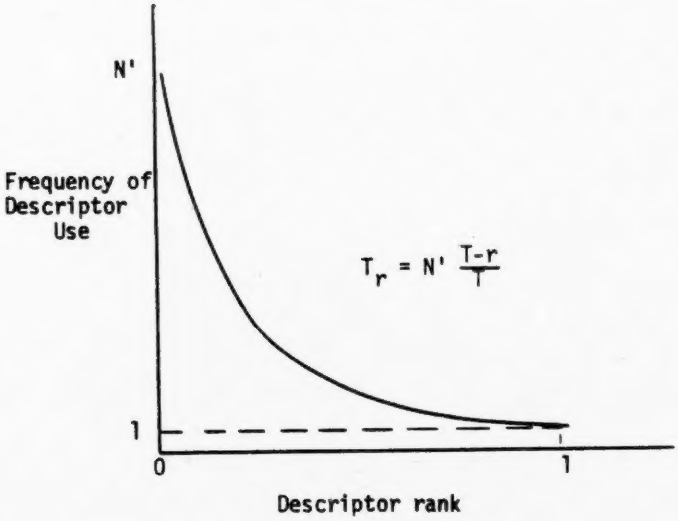


Fig. 4.2 -- Normalized term usage vs. rank

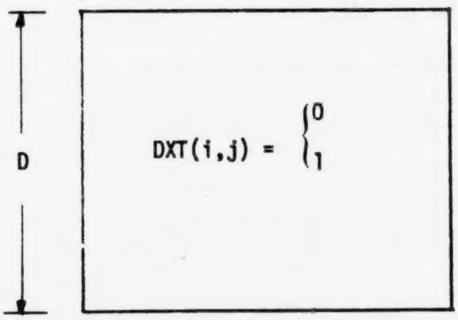

T
D
$DXT(i,j) = \begin{cases} 0 \\ 1 \end{cases}$

Fig. 4.3 -- Document-term association matrix

$$R_q = T_o\left(\frac{T_o}{J}\right)^{n-1}$$

The last relationship given is also extended to disjunctive combinations of terms by noting that, "for r total documents equal to the sum of all 'or' terms, the expected number of different documents out of J documents will be,"

$$(r)_{\text{adjusted}} = J(1 - \frac{1}{J})^r$$

In going through the above derivations it becomes clear that terms are assumed to be independently assigned, and that term co-occurrences are independently distributed. Consequently the Raver estimations do diverge greatly from actuality. However, this derivation is unique in that the term frequency of use distribution is, albeit implicitly, assumed to be of some standard form represented by a stable class of functions -- in this case the log function. This notion, as well as that of the need to explicitly normalize the term-frequency of use versus rank distributions, is used in the proposed methodology in this report.

A different perspective is taken by King and Bryant (23) who deal with the issue of quantity output in the context of an overall system evaluation scheme, in which relative frequency of indexing consistency (aggregated over the thesaurus, indexers and corpus) at a point in time is determinable. As such, the expected number of documents is simply the number of documents in the file that are relevant to the inquiry. That is, if a K term inquiry were submitted with the conjunctive requirement that the retrieved documents be described by all K terms, then the expected number of documents retrieved would be:

$$R_q = D \sum_{n_2,n_3=1}^{K} P_2^{n_2} P_3^{n_3} Q_{n_3}$$

$$\text{s.t. } n_2+n_3=K$$

where Q_{n_3} = the portion of the corpus that "should" contain n_3 of the K terms in the inquiry

P_2 = the relative frequency of a document being indexed when it should not be (a Type II error)

P_3 = the relative frequency of a document being indexed when it should be indexed by the inquiry terms

D = the number of documents in the corpus.

The above analysis is basically dependent on the assumption of independence of indexing errors. That is to say, if the assignment of terms to documents is sufficiently consistent, a norm can be observed about which statistical fluctuations will sum to zero -- if the indexing errors are indeed independently distributed. A second assumption is that there exists the ability to determine the fraction of the corpus that should contain (be indexed by) the terms in the request. Neither of these assumptions seems operationally practical, and is rather an awkward basis for determining R_q. Clearly, one of the desirable attributes of an operational R_q estimation process is that it not require unwieldy computations, or data not readily available.

Another somewhat different scheme, which indirectly addresses the issue of quantity output, is investigated by Shumway (113). This procedure involves an estimate of the total number of relevant documents in a corpus, through the use of sampling techniques common to probit analysis, and then estimating, with appropriate confidence intervals, the number of documents necessary to output in order to retrieve a certain specified quantity of the relevant documents in the corpus.

The estimation process entails taking an initial sample for which the recall ratio (see Fig. 3.1) is determined. Then a second sample (of the same size) is taken, and based on the overlap of common "relevant" documents an estimate of the total set of relevant documents in the corpus is made. This technique involves the use of the hypergeometric distribution, and requires that the samples be random.* The result of the sample sequence is used to construct a search characteristic curve which measures or reflects the number of documents needed to be retrieved in order to get a certain number of relevant documents.

Wiederkehr (148) also utilizes the search characteristic curve to estimate quantity output, and presents the interesting notion that any search strategy has an equivalent series of single stage random searches to generate the desired number of relevant documents in the corpus. The notion of defining a search inquiry as a multiple of single stage random searches is very useful, and will be incorporated in the proposed methodology discussed in the next section of this chapter.

The usefulness of a search characteristic curve is limited by the requirements for data sampling, and the judgment consistency of what is or is not relevant to an arbitrary inquiry. Also, the distribution characteristics essential to the probit/hypergeometric are not sufficiently satisfied by a DRS.

---

*For a more complete discussion of this procedure see Feller (50).

In summary, the principal efforts to date have not developed an operational procedure for estimating R_q that could be useful to a manager or designer of a DRS. The common assumption of random assignments of descriptors to documents or its equivalent term-term independency assumption is not satisfied by actual DRSs. In addition, those procedures that could, albeit indirectly, lead to estimates of R_q require an impractical amount of data and extensive relevance judgments.

4.3 PROPOSED METHODOLOGY FOR DEVELOPING THE R_q MEASURE

As noted, previous attempts to construct a retrieval quantity measure have, in general, failed to correctly represent the characteristics of DRS components, and also have not taken advantage of the statistical regularity common to certain components of coordinate indexed DRSs.

At the onset of developing an operational tool, it is advantageous to indicate the desirable characteristics that the measure should possess. Four such characteristics are:

1. The R_q measure should be defined in terms of the basic DRS components (or their equivalent distributions).
2. The measure should use data that is convenient to obtain in operational DRS settings, and easy to construct for those systems in the design stage.
3. The value of the measure should be easy to compute.
4. The measure should possess stability to allow (in the dimension it measures) -- (a) monitoring of intra-system changes, and (b) inter-system comparisons.

As a preamble to the specifications of the R_q measure, a brief review of the basic DRS components, relationships and characteristics will be given. Where a DRS characteristic or relationship is recommended for incorporation in the measure, a hypothesis will be made about the particular system property. In Chapter 5, the various hypotheses stated in this chapter will be analyzed for acceptance or rejection.

4.3.1 Fundamental DRS Relationships

As noted in an earlier chapter, the basic DRS components are:

(a) the system corpus -- D

(b) the system thesaurus -- T

(c) the term-document distribution -- DXT

The DXT distribution is the basis from which all other DRS characteristics are derived. For the class of DRSs of interest to this analysis the DXT matrix is binary, and a hypothetic example is given in Fig. 4.3.

If one arrays the columns in the DXT matrix such that the term with the greatest frequency of use is given rank 1, and the second most frequently used term given rank 2, and so on, the resulting DXT matrix can be represented by the term frequency of use distribution in Fig. 4.2. Note that the most frequently used descriptor is assigned N_{max}, as the highest frequency of use, and the least used (>0) descriptor N_{min}. When $N_{min} = 1$, the frequency-rank distribution is effectively normalized. However, if $N_{min} > 1$, as is the case in certain truncated distributions, the distribution can be normalized by the division of N_{min}, as indicated in Fig. 4.2.

The term frequency of use distribution is a commonly available DRS statistic, and a preferred data source for the R_q estimation process. In order to formally describe the term frequency of use distribution two hypotheses will be offered:

I. The term-frequency-of-use versus rank distribution is a decreasing concave (convex) function in ordinal (log-log) space, and is closely approximated by the Mandelbrot-Estoup-Zipf (MEZ) distribution.

The Mandelbrot-Estoup-Zipf* (94, 95, 96, 153) relationship is defined for the distribution of word frequency in an unrestricted language in which the relative probability of occurrence of a word or term is defined to be

$$P_{r_i} = K(r_i + B)^{-\alpha}$$

where R_i = the rank of term i that is used N_i times

P_{r_i} = probability of occurrence of the term i (with rank r_i)

$K = e^{-\beta,t_0}$; derived from the exponential law for optimum codes; for this application $e^{-\beta,t_0}$ is a constant to be determined empirically

$\alpha = \beta_1/\beta$; also a constant to be determined empirically.

The basic form of the MEZ canonical form is illustrated in log-log space in Fig. 4.4. For comparison, the more specific Zipf's Law (a special case of the MEZ form) is also indicated.

The term-frequency of use distribution is a representation of the column marginals of the DXT distribution. Taking the row marginals of

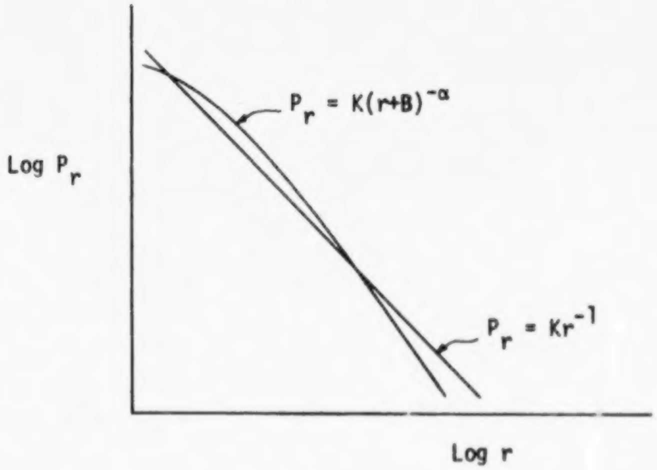


Fig. 4.4 -- Term frequency of occurrence versus rank in log-log space

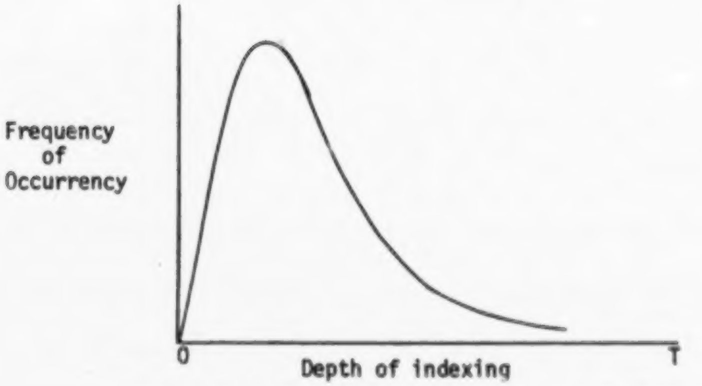


Fig. 4.5 -- Typical depth of indexing distribution

the DXT matrix yields the depth of indexing distribution. A typical depth of indexing distribution is illustrated in Fig. 4.5. This distribution displays the assignment of terms to documents, and will be referred to again in a later chapter as a source of data to define the degree of homogeneity of a DRS corpus.

An additional DRS characteristic, and one of central relevance to the R_q measure, is the term-term (TXT) correlation matrix. This matrix is defined as follows:

$$(DXT)^T DXT = TXT$$

The element TXT(i,j) represents the number of co-occurrences of term i and j. For example, if term i and j were assigned to the same n documents, the value of TXT(i,j) would be n. Another way of defining the elements of TXT is that they are the inner product of the i^{th} column vector of DXT with the j^{th} column vector of DXT. Also the matrix TXT is a symmetric distribution.

Having defined the TXT distribution, the second hypothesis about the term frequency of use distribution can be made:

II. The term-term co-occurrence distribution is not generated by a process which selects terms for assignment independent of one another.

Since the R_q measure emphasizes quantity, a third hypothesis is of interest, and is also based on the data in the TXT distribution:

III. Terms with the same frequency of use have essentially the same statistical characteristics in the TXT distribution.

4.3.2 Inquiry Definition and Generation

The process of inquiry generation is initiated by the system user, who upon "experiencing" a need for information, converts that need into a "natural language" request, and then interprets (with or without the aid of the DRS personnel) the request into a formal DRS inquiry. A formal inquiry is defined as consisting of terms from the system thesaurus that are coordinated in accordance with the system grammar. The rules of coordination to be used in this analysis are defined by Boolean Algebra. The explicit operations used for term coordination are: Union or logical sum $(+)$, Intersection or logical product $(\cdot)$, and exclusion or logical negation $(-)$.

The pertinent characteristics of the inquiry are the form (the number of terms and operators by type) and the frequency of use of the terms. The semantic characteristics of the inquiry are not used in the R_q determination, as it is assumed that terms with the same frequency of occurrence have essentially the same term-term co-occurrence characteristics (Hypothesis III above). This assumption, which is proven in the next chapter, simplifies the inquiry generation process for developing hypothetic DRS workloads for DRSs in the design stage.

4.3.3 Inquiry -- Retrieval Quantity Measure Relationship

The basic variables relating inquiry terms to documents retrieved are:

(1) term-frequency of use $(f(i))$

(2) term-term co-occurrence values $(TXT(i,j))$

For the logical operators of ".", "+", and "-", the following relationships hold for elementary two-term inquiries:

Request	Inquiry	Output Quantity
T_i	i	f(i)
T_i and T_j	i·j	TXT(i,j)
T_i or T_j	i+j	f(i)+f(j) - TXT(i,j)
T_i and not T_j	i-j	f(i) - TXT(i,j)

Therefore, for all elementary two-term inquiries, knowledge of the term frequencies of use and their co-occurrence value is sufficient to determine the output quantity. For more complex inquiries in which many terms are coordinated the determination of R_q is not so simple. It follows, however, that if the single terms in the above example were replaced by groups of, say, conjunctively related terms, the same relationships would hold. For example, given groups E and F with a logical product 0_{EF}, the following is true:

Inquiry	Output Quantity
E·F	0_{EF}
E+F	$0_E+0_F-0_{EF}$
E-F	$0_E - 0_{EF}$

The above relationships hold for sequences of disjunctively related groups of conjunctively coordinated terms, or for conjunctively related sequences of disjunctively coordinated terms. It can be shown that any retrieval specification (in the propositional or predicate calculus) on the set of thesaurus terms can be represented in disjunctive or conjunctive normal form. A disjunctive normal form is a disjunction of clauses with no repetition of terms within the clauses. A clause is simply a finite conjunction of terms (where negation is defined as a negative conjunction). Also every disjunctive normal form has a dual conjunctive normal form (53, 103, 108, 109). Thus, no matter how complex, an inquiry can be converted to a string of clauses that can be evaluated for quantity output, as per the relationships in the above example. The crucial value to determine is the logical product.

4.4 HYPOTHESES FOR RETRIEVAL QUANTITY ESTIMATIONS

Given the search strategy of direct match, two methods of estimating the logical product of inquiry terms, and the value of R_q are discussed in this section.

The problem of determining R_q for a multiterm inquiry is illustrated by the following example. For an n term disjunctively coordinated inquiry, $T_1+T_2+\ldots+T_n$, the estimate of R_q is:

$$R_q = f(1) + f(2) + \ldots + f(n) - \text{Logical Product}_{(1,\ldots,n)}$$

The simplest model for estimating the logical product of two or more terms is one that assumes that the descriptor assignment to a document is a random assignment. This case has been noted as being basically incorrect; however, it can be employed as a stepping stone to an eventual solution. For this model, the logical product of two or more terms is:

$$\text{Logical Product}_{(i,\ldots,n)} = \frac{f(i)\cdot f(j)\ldots f(n)}{D^{n-1}}$$

and, R'_q for a n term disjunctive inquiry, $T_i+T_j+\ldots+T_n$ is:

$$R'_q = f(i) + f(j) + \ldots + f(n) - \frac{f(i)\cdot f(j)\ldots f(n)}{D^{n-1}}$$

It can be shown* that the actual value of the logical product and the "random-case" values do diverge significantly. However, if one makes the hypothesis:

IV. There exists a stable statistical relationship between the actual term-term distribution and the hypothetical "random case" distribution,

then the above formulation yielding R'_q can be modified to yield an accurate estimator of R_q. From the above hypothesis the proposed modification is:

$$R_q = \gamma R'_q$$

or what is equivalent

$$\text{Logical Product}_{(1,2,\ldots,n)} = \gamma_{1,2,\ldots,n}\left(\frac{f(1)\cdot f(2)\ldots f(n)}{D^{n-1}}\right)$$

This hypothesis will be tested for acceptance or rejection in Chapter 5. If the hypothesis is accepted then a very convenient method for estimating R_q will be available.

---

*A statistical test of an actual DRS is performed in Chapter 5 to demonstrate that the distribution of logical products of terms is not equivalent to a "random-distribution."

Given that the proportion γ proves to be acceptable, the proposed utilization of the proportion for multi-term inquiries is illustrated in the following example:

Inquiry: $T_1 \cdot T_2 \cdot T_3 \cdot T_4$

Estimation of R_q:

(1) Let $f(1') = \gamma_{1,2}\left(\frac{f(1) \cdot f(2)}{D}\right)$

(2) Let $f(2') = \gamma_{1',2}\left(\frac{f(1') \cdot f(3)}{D}\right)$

(3) Therefore, $R_q = \gamma_{2',4}\left(\frac{f(2') \cdot f(4)}{D}\right)$

A second model for estimating the logical product of two or more terms can be constructed by using the Row (MR) marginals and column (CR) marginals, and the total sum (TS) of marginals for the TXT matrix. For this model, the expected value of the logical product of two or more terms is:

$$\text{Logical Product}_{(1,2,\ldots,n)} = \sum_{i,j}^{n} \frac{MR_i \cdot MC_j}{TS}$$

where MR_i = sum of the term co-occurrences in Row i -- for term i with terms $1,\ldots,T$

MC_j = sum of the term co-occurrences in column j -- for term j with terms $1,\ldots,T$

$$TS = \sum_{k=1}^{T} MR_k = \sum_{k=1}^{T} MC_k$$

and, R''_q for an n term disjunctively coordinated inquiry, $T_i + T_j + \ldots + T_n$ is:

$$R''_q = f(i) + f(j) + \ldots + f(n) - \sum \frac{(MR_i)(MC_i)}{TS}$$

where an analogous hypothesis, to model 1, is

$$R_q = \lambda R''_q$$

or what is equivalent

$$\text{Logical Product}_{(1,2,\ldots,n)} = \lambda_{1,2,\ldots,n}\left[\Sigma \frac{MR_i \cdot MC_j}{TS}\right]$$

Using experimental data in Chapter 5, the above relationships will be tested to determine if they can be accepted or rejected for use as an operational tool.

> All the business of life, is to endeavor to find out what you don't know by what you do
>
> The Duke of Wellington

Chapter 5

THE RETRIEVAL QUANTITY MEASURE: EXPERIMENTS AND RESULTS

5.1 INTRODUCTION

The purpose of this chapter is to analyze the various hypotheses made, thus far in this report, about the fundamental characteristics and relationships of coordinate-index DRSs, and to construct and test an operational R_q estimation model for systems that are established or in the design stage.

In the preceding chapter the following hypotheses about DRSs were stated:

(1) the term-frequency-of-use versus term rank distribution is a monotonically decreasing concave function in log-log space, and is closely approximated by the M-E-Z canonical form.

(2) the term-term co-occurrence distribution is not generated by a process which selects terms for assignment independent of one another; that is to say, the term co-occurrence distribution is not the result of random sampling from the thesaurus.

(3) the co-occurrence value of two terms is directly proportional to a function of the frequencies of use for the terms, and can be predicted as a function of that factor.

(4) terms with the same frequency of use have essentially the same statistical characteristics. That is, two terms i and

j, with frequencies of use f(i) = f(j) will have approximately the same number of co-occurrences with other terms in the thesaurus.

(5) the Retrieval Quantity (R_q) of a coordinate index DRS can be predicted for formal inquiries.

One of the principle aims of this chapter is the analysis of these hypotheses for acceptance or rejection. The required experiments and analyses for this task and for the construction of the R_q model are discussed next.

5.2 EXPERIMENTS: SETTING AND DESCRIPTION

Experiments for the analysis of the above hypotheses were performed at the Institute of Library Research Information Processing Laboratory at the University of California, Berkeley, California. At the time of the experiments, the Laboratory facilities consisted of three Sanders CRT-remote on-line terminals to a IBM 360, Model 40, 128K system. The CRTs had keyboard input and visual display output, and were capable of simultaneous operation.

The Laboratory system was equipped with three search grammars, and eight word association files (including direct match search capability).

The experiments were set to take place over a period of time in which the Laboratory DRS corpus and thesaurus were expanded. The original plan called for a three-stage growth sequence, but only the first and second stages were realized. The system characteristics for the two stages are tabulated in Table 5.1, and the term-freqaency of

Table 5.1

ILR DOCUMENT RETRIEVAL SYSTEM

Characteristics	Stage 1	Stage 2
Corpus (documents)	≈ 300	400
Thesaurus (terms)	368 (348 active)	393 (375 active)
Average depth of indexing	14	12-13
Average term usage	3-4	3-4

Table 5.2

DATA BASE SAMPLE

Characteristics	
Corpus	102
Thesaurus	≈320 (307 active)
Average depth of indexing	14
Average term usage	3-4

use versus term rank distribution for the system, at the end of stage two, is shown in Fig. 5.1. Samples of the system thesaurus and term-document assignments are included in Appendix B. The DRS corpus is composed exclusively of documents on information science, and can be appropriately classified as being homogeneous. For a more complete description of the Laboratory and its research projects see Maron, et al. (98).

5.2.1 Experiments and Analysis

The data collection and analysis involved several steps. The first consisted of gathering of the DRS responses, over the two stages of system growth, for a set of formal inquiries. The second step entailed an analysis of a data sample from the DRS term-document distribution[*], and the third, the evaluation of the retrieval quantity model. In the next two sections, 5.3 and 5.4, all these steps are discussed in detail, and the hypotheses are analyzed for acceptance or rejection.

5.3 DOCUMENT RETRIEVAL SYSTEMS -- COMMON CHARACTERISTICS

In this section, the issues of statistical regularity among coordinate indexed DRSs, and the data analysis which demonstrates the statistical similarity of the test system to other DRSs, of different size and subject matter, are discussed.

A number of researchers, Brookes (20), Fairthorne (49), Mandelbrot (94, 95, 96), to mention a few, have observed that there are certain statistical regularities common to a variety of documentation systems and activities. Fairthorne (49), in fact, presents a brief survey of

---

[*] See Appendix C for a description of the data sample.

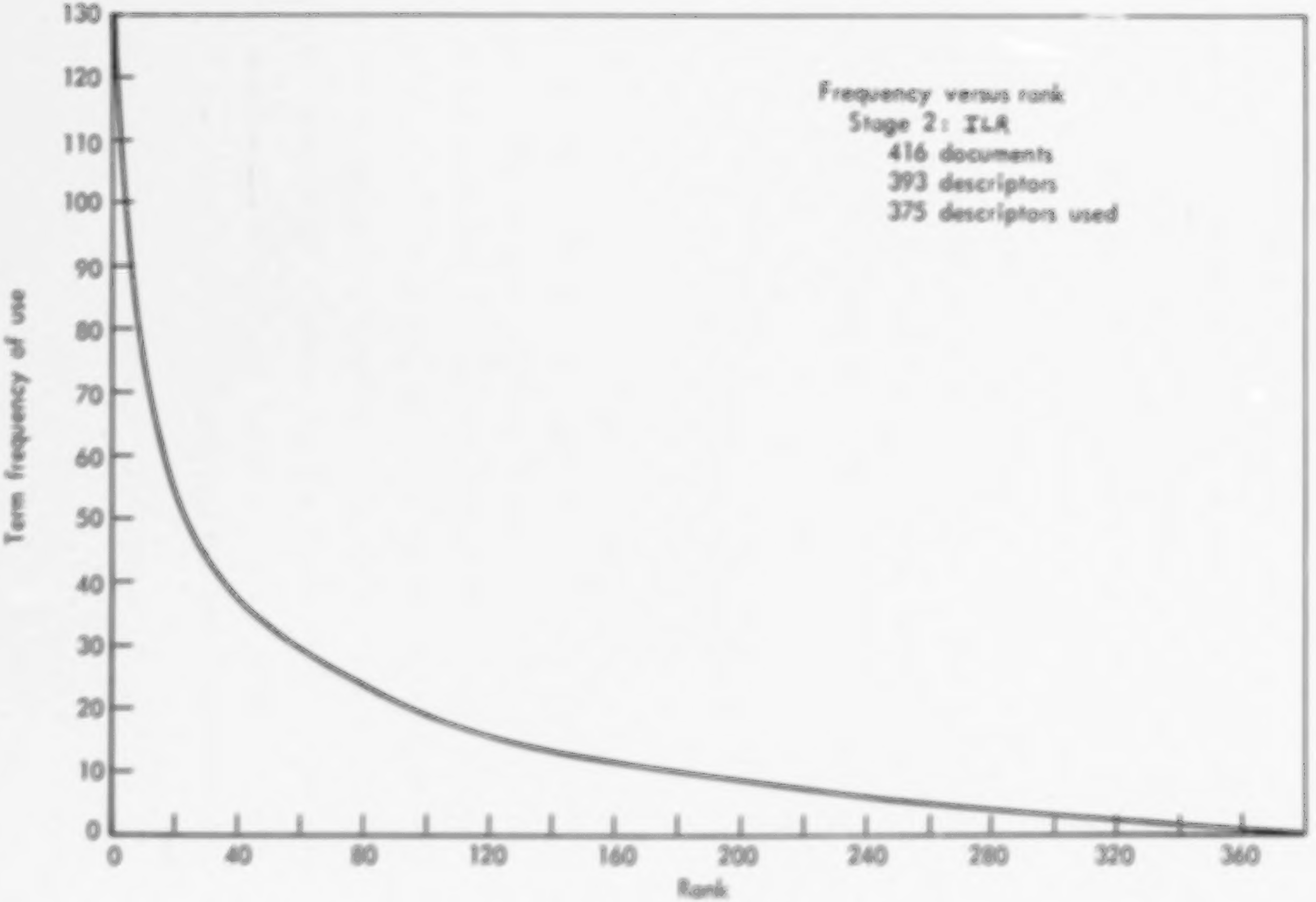


Fig. 5.1 — ILR document retrieval system

this topic.[*] All of these findings revolve around the concept that the underlying behavior of DRSs is "hyperbolic" in nature (49).

Of interest to this analysis are the characteristics of derived-manipulative indexed DRSs that exhibit similar properties, independent of systems size and subject matter. The basic relationship for DRSs is the index term -- document distribution, from which all the term-term, and document-document functional relationships can be derived. Therefore, if the term-document distributions of different DRSs can be shown to be statistically similar, or definable by an analytic/canonical form, the argument for statistical regularity among DRSs can be accepted. The principal vehicle for showing this is the term-frequency-of-use distribution.

5.3.1 The Term-Frequency-of-Use Distribution

The preferred characteristic to use to determine if there is a statistical similarity among DRSs is the term-document (TXD) distribution. However, the TXD distribution is awkward to deal with and is rarely ever published. Thus the strategy taken is to use surrogate distributions; namely, the term-frequency-of-use versus term rank, the term usage versus the cumulative frequency distribution, and the depth of indexing distribution. The first two distributions, in particular, are readily available from published research and all three distributions are convenient to illustrate. The relationships between these distributions and the TXD matrix are illustrated in Fig. 5.2.

---

[*] A richer but unfortunately abstruse discussion is given by Mandelbrot (94-96).

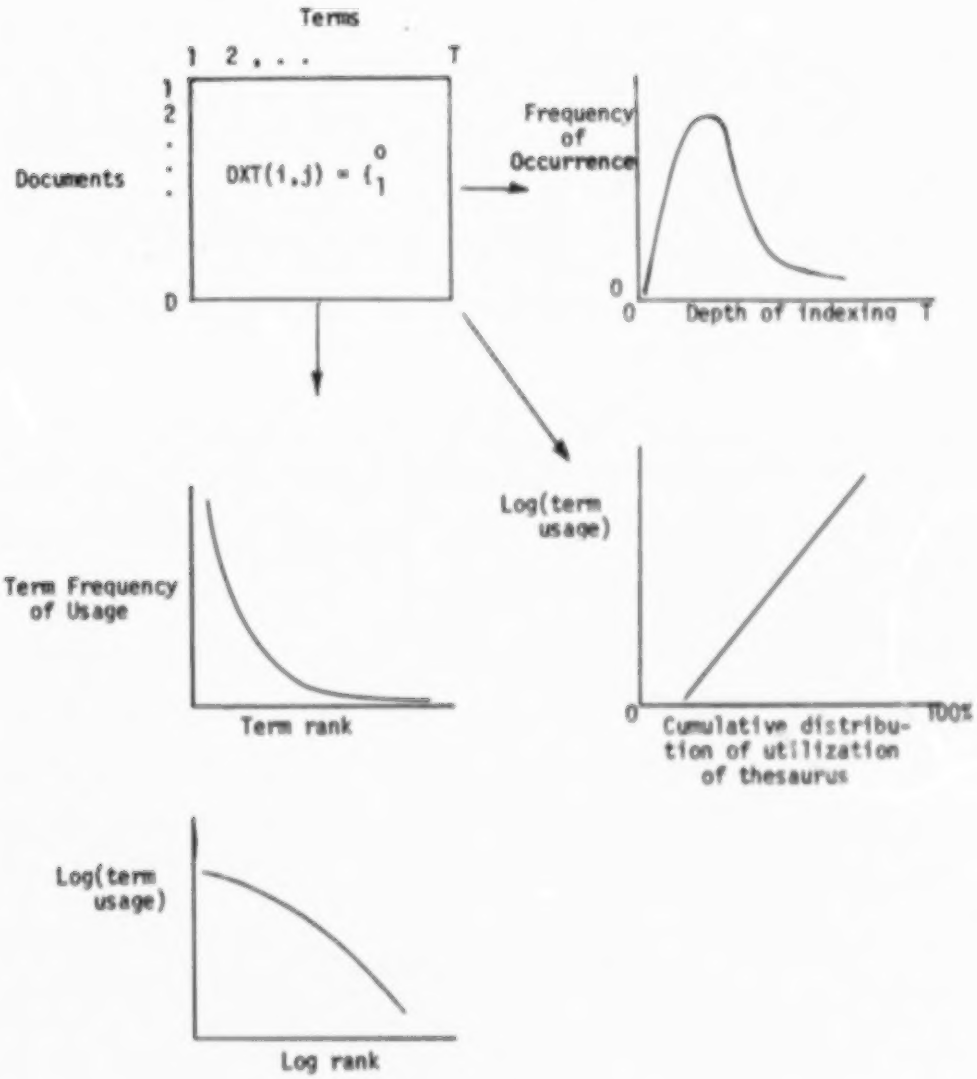

Terms
1 2 , . . T
1
2
Documents
D
DXT(i,j) = {0, 1}
Frequency of Occurrence
0
0
Depth of indexing
T
Term Frequency of Usage
Term rank
Log(term usage)
0
Cumulative distribution of utilization of thesaurus
100%
Log(term usage)
Log rank

Fig. 5.2 -- Illustration of relationships between the term document matrix and the term frequency of use vs. rank distribution, and the term usage vs. cumulative usage distribution, and the depth of indexing distribution

Figure 5.3 shows in log-log space the term frequency distribution for the test system sample, the test system, and the three larger DRSs investigated by Litofsky (90). All the curves are concave monotonically decreasing relationships. The two DRSs investigated by A. D. Little (1) are shown in Fig. 5.4, and these systems also display the same concave monotonically decreasing term frequency of use versus rank in log-log space. It is important to note that these systems are terrifically different in size, and have different subjects for corpus content.

In addition, Houston and Wall (68) and Wall (143) have analyzed some 14 DRSs and plotted their term-frequency of use versus the cumulative percent of thesaurus utilization.* Their plots are reproduced in Figs. 5.5 and 5.6. All the systems plotted exhibit a remarkable linearity for the postings per term versus the cumulative distribution, which lead Houston and Wall to conclude that the number of terms T in a system vocabulary varies directly with the log of TU, the total number of term uses, and has the form:

$$T = a \operatorname{Log}_{10}(TU + b) - c$$

where $a = 3300$

$b = 10000$

$c = 12600$

for values of TU between 10,000 and 1,000,000. As further evidence of statistical regularity, the three systems analyzed by Litofsky (90) and the ILR systems are plotted in the Houston-Wall dimensions. These

---

*Fairthorne (49) points out that the two methods of illustration are just different ways of showing the same characteristics.

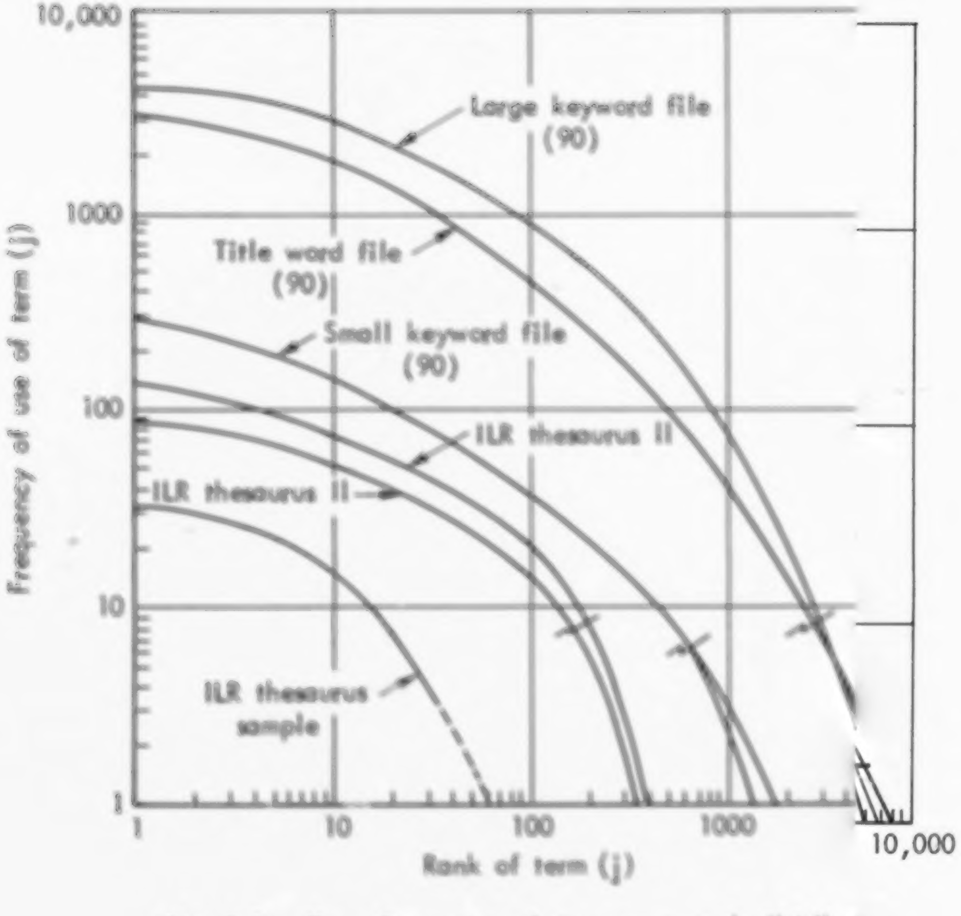


Fig. 5.3—Term frequency of use versus rank distribution

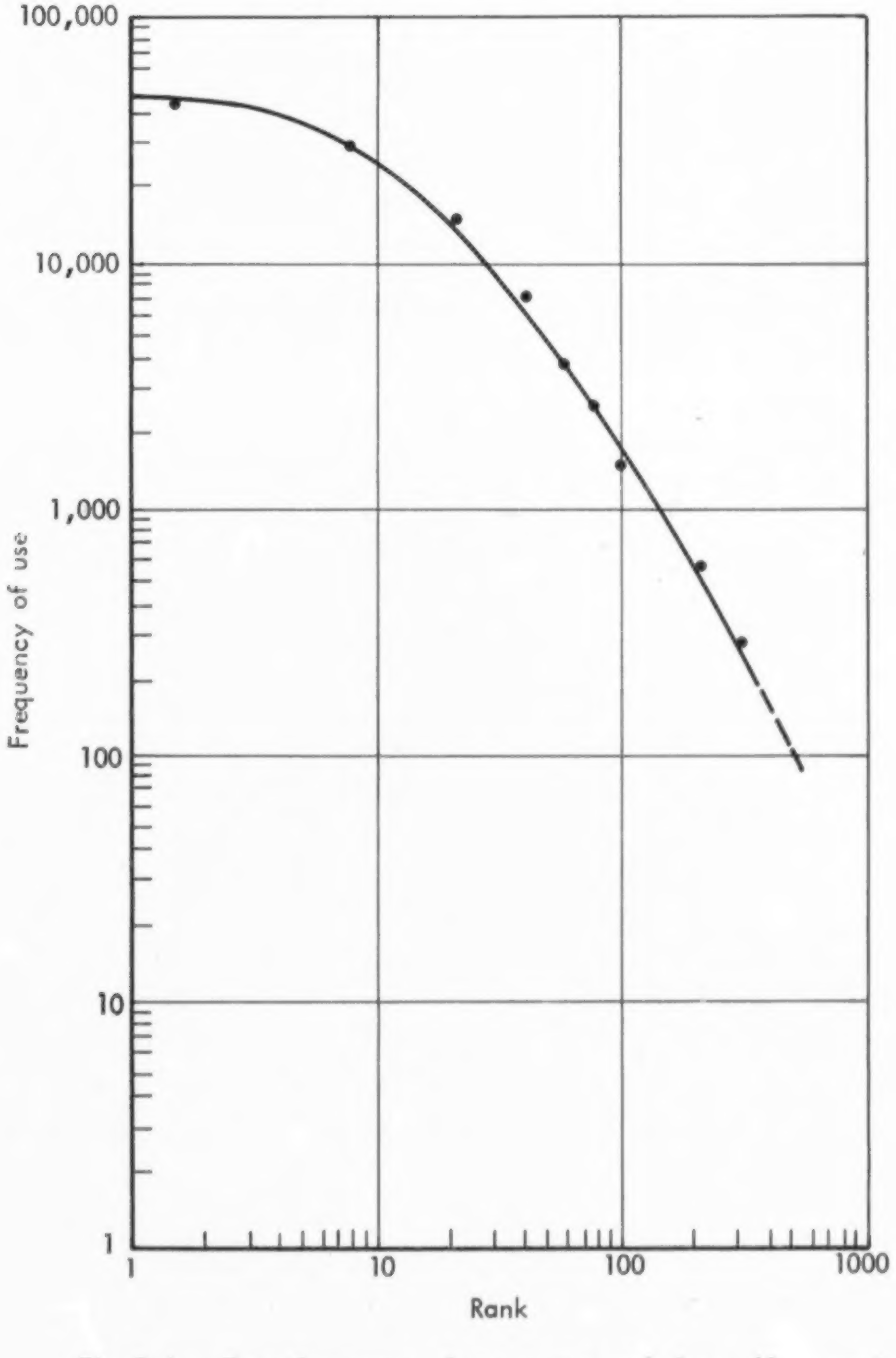


Fig.5.4 — Term frequency of use versus rank for a 10 percent sample of the Industrial Collection System investigated by A.D. Little (1, 2)

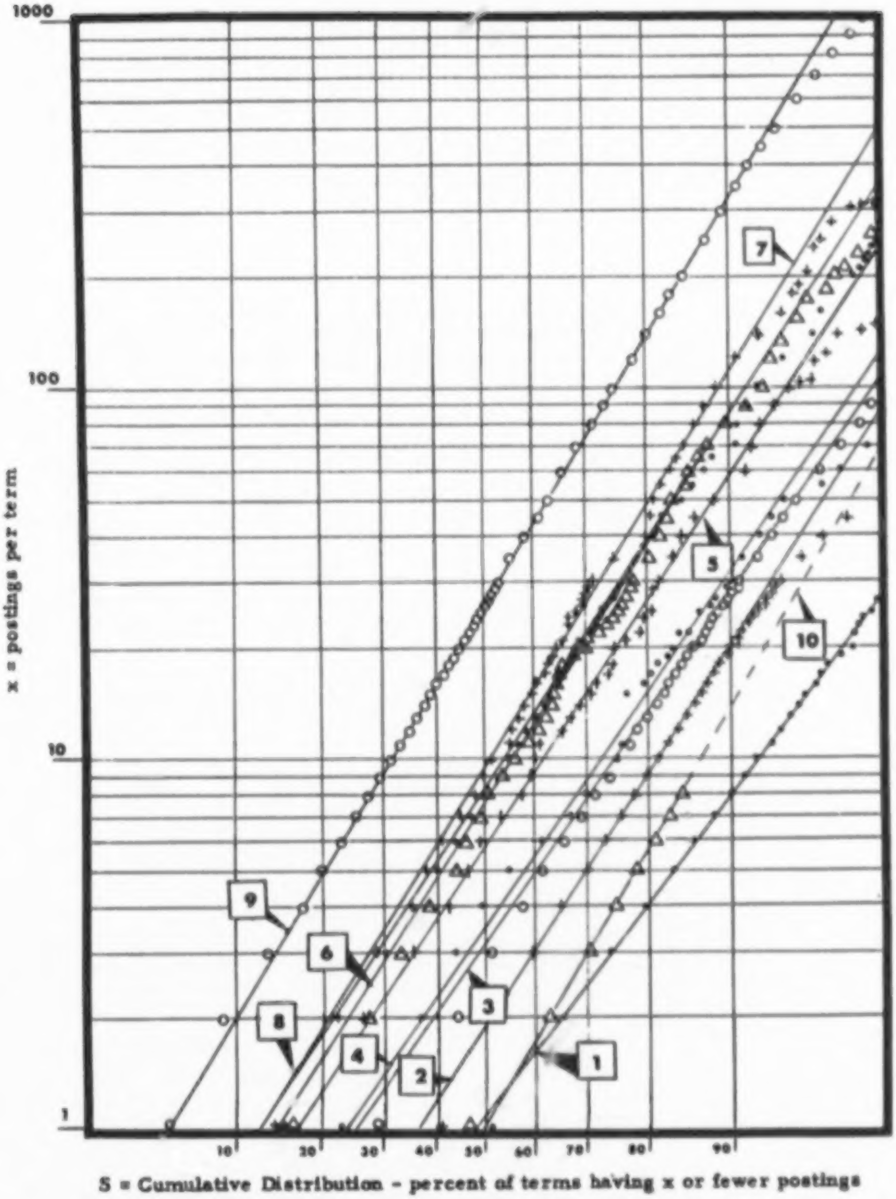


Fig. 5.5 -- Term usage versus cumulative thesaurus utilizations of thesaurus for systems investigated by Houston and Wall (68). (See Table 5.3 for systems corresponding to numbered curves)

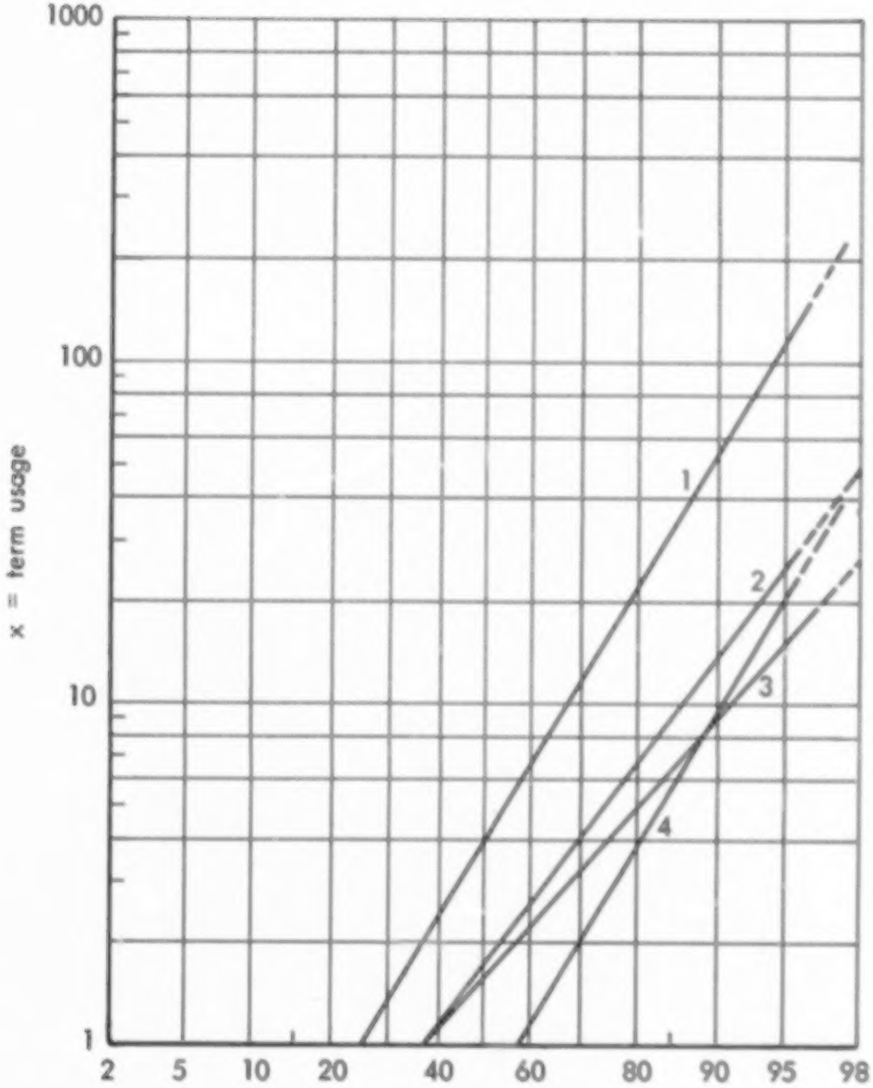


Fig.5.6 — Term usage versus cumulative utilization of thesaurus for systems investigated by Wall (143) (see Table 5.3 for systems corresponding to numbered curves)

plots are also linear and are shown in Figs. 5.7 and 5.8. The above relationship holds quite nicely for the keyword files analyzed by Litofsky. The ILR system, however, is too small as its TU is < 10,000, and the above constants require adjustment; the form of the relationship, however, is satisfied.

This empirical evidence is even more impressive when one compares the range in corpus and thesaurus size, the different subjects covered, and the variation in index term utilization. These pertinent system characteristics are tabulated in Table 5.3.

5.3.2 The Term-Frequency-of-Use Canonical Form

In addition to the graphical interpretation, which implies strong statistical stability, a number of efforts have been made to define the term-frequency-of-use versus rank relationship analytically.

The most well-known attempt to define in equation form a general relationship between term frequency of occurrence and term rank is by Zipf (152), who suggested the form:

$$f(r) \cdot r = K$$

where K = a constant for a particular (large) sample of text in any language

$f(r)$ = the frequency of occurrence of the term with rank r

r = term rank; a positive integer.

This expression is based on empirical observation of free or running text, and as noted by Mandelbrot (95) and Fairthorne (49), it is an extension of the earlier work of J. B. Estoup in 1916 and J. Willis in 1922. Mandelbrot (94, 95) using communication or information theory

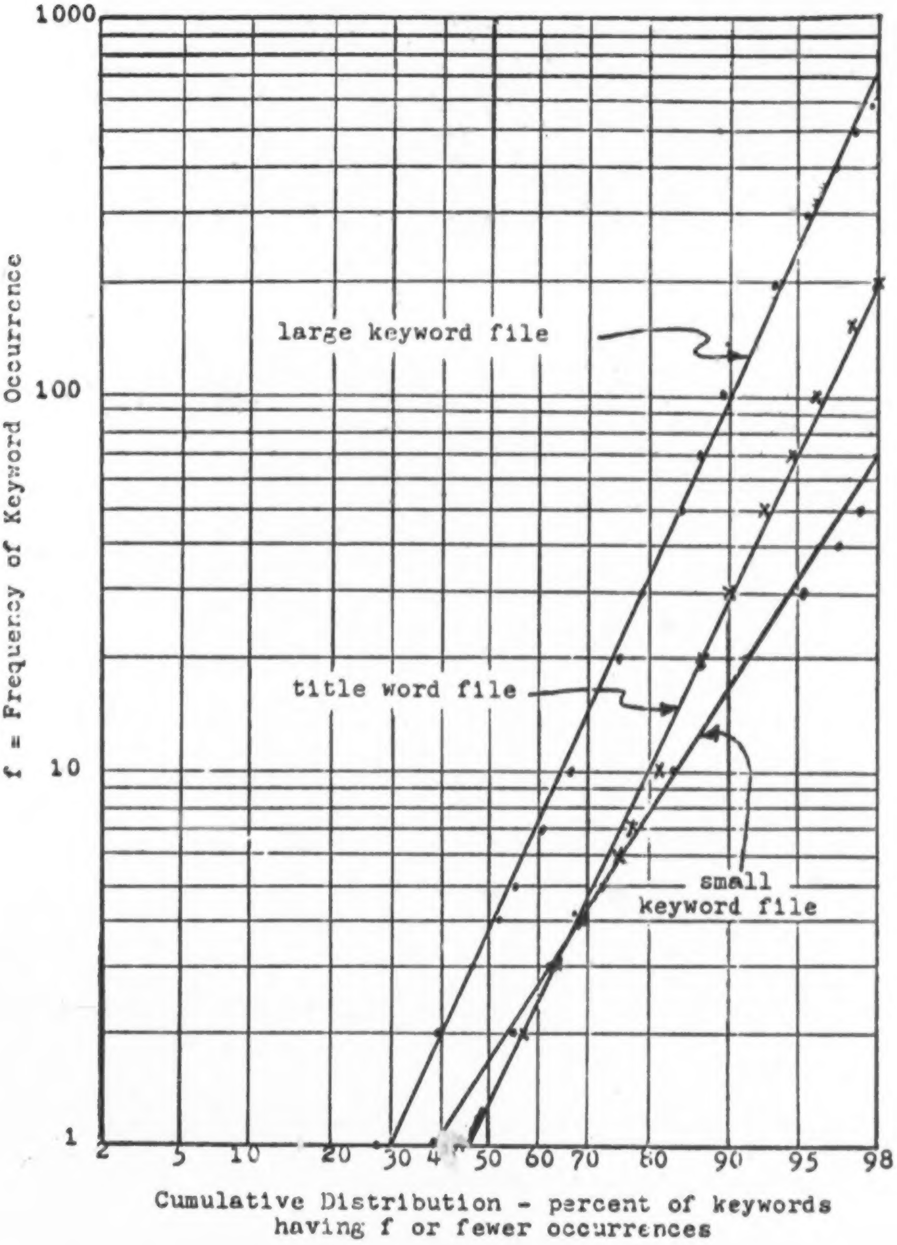


Figure 5-7
Log-Probability Plot of Keyword Distribution (90)

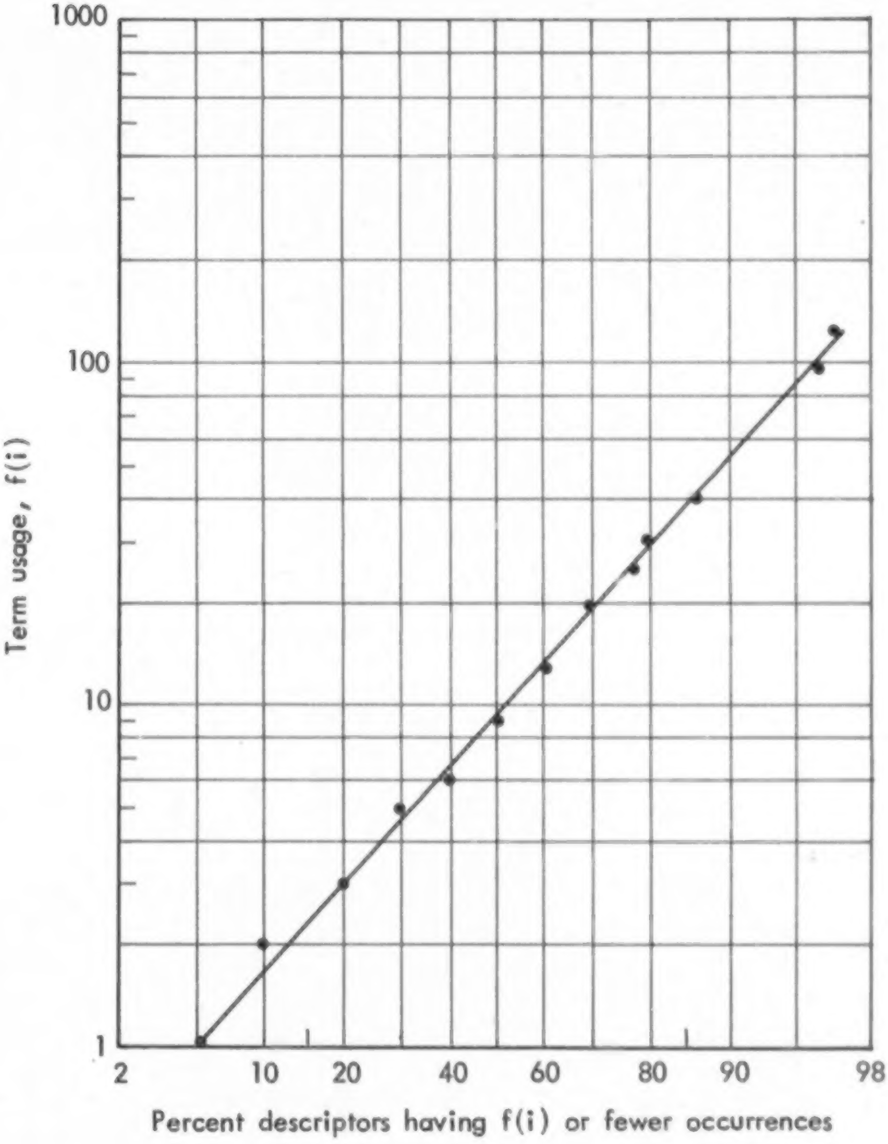


Fig.5.8 — Log probability of descriptor usage ILR test system

Table 5.3

SYSTEM CHARACTERISTICS

System Characteristics	ILR	Litofsky (90)			A.D. Little (1,2)			Houston and Wall (68)[a]						
		Small Keyword	Large Keyword	Title Word	ASTIA	AEC	Industrial	1	2	3	4	5	6	7
Corpus size	400	2254	46821	46942	70000	N.A.	5000	303	2100	2992	1468	3253	5087	6930
Thesaurus size	390	2557	8044	13309	7145	12704	4780	1108	2100	3459	4060	4890	5280	5410
Average terms per document	14	9	10	6	N.A.	N.A.	13	14.3	8.5	15.0	3212	32.3	31.8	31.3
Average documents per term	3	8	58	21	N.A.	N.A.	N.A.	3.9	8.5	13.0	11.7	21.4	30.6	40.0
Number of terms used once		992	2189	5879	N.A.	N.A.	N.A.	—	—	—	—	—	—	—

[a] 1 Private Experimental Index - 1960
2 Dupont Engineering Department - 1956
3 AFOSR Project ECHO - 1962
4 Uniterm Index to Chemical Patents 1961
5 Uniterm Index - 1961
6 Uniterm Index - 1961
7 Uniterm Index - 1961

Table 5.3--Continued

System	Houston and Wall (68)[a]			Wall (143)[b]			
Characteristics	8	9	10	1	2	3	4
Corpus size	9018	195000	2092	--	2356	3569	--
Thesaurus size	7730	6690	15595	19106	1755	3370	8087
Average terms per document	27.6	5.2	65.5	—	—	—	—
Average documents per term	32.2	148.0	8.8	—	—	—	—
Number of terms used once	—	—	—	—	—	—	—

[a] 8 Uniterm Index - 1961
9 DDC - 1960
10 Private Research Report Index - 1959

[b] 1 KWIC Index (IBM Corp.)
2 Published Uniterm Index
3 Subject Heading Index - NASA
4 KWIC Indexes to IBM and ASTIA Reports

as a basis has derived a relationship, between word frequency of use and the rank of a word, that is more general than Zipf's, and of which Zipf's is a special case. Because of the various contributors, this relationship will be referred to as the Mandelbrot-Estoup-Zipf (MEZ) distribution, and has the form:

$$f(r) = K(r+B)^{-\alpha}$$

For $B=0$ and $\alpha=1$, the above relationship reduces to Zipf's "Law". However, Zipf's equation calls for a linear plot of slope minus one in log-log space, which is not satisfied (even with congruent intercepts) by the curves plotted in Figs. 5.3 and 5.4.

For the purposes of this analysis it will be sufficient to show that the MEZ canonical form is close to the actual term-frequency of use versus term rank distribution. To illustrate how the parameters K, B and α are defined for a DRS (at a certain point in time), the test system characteristics will be used. For the test DRS:

$D = 102$

$T = 370$

$T' = 307$ (the number of active terms in the thesaurus)

$\hat{D} = 14$ (the average depth of indexing)

$f(r=1) = 32$ (the frequency of use of the term with rank = 1)

$f(r=300) = 1$ (the frequency of use of a term with rank $\overset{\sim}{\wedge}300$)

Zipf [see Booth (10)] has noted that a term will occur once if

$$1.5 > \hat{T}\, P(r) \geq 0.5$$

where P(r) = the probability of occurrence of a term with rank r

$$= \frac{f(r)}{\sum_{r=1}^{T'} f(r)}$$

$\hat{T}$ = the total number of term occurrences

$$= \sum_{r=1}^{T'} f(r)$$

The above relationship can be generalized for a term occurring n times,

$$(n + 1/2) > \hat{T}\, P(n) \geq (n - 1/2).$$

Substituting the MEZ form for P(n) yields

$$(n + 1/2) > \hat{T}\, K'(r + B)^{-\alpha} \geq (n - 1/2).$$

For a term with the highest rank, $r_{max} \approx T'$, and where $B < T'$ (which is always the case--see Mandelbrot (95)), and $n = f(T') = 1$, the inequality becomes:

$$1.5 > \hat{T}\, K'(T')^{-\alpha} \geq .5$$

Because the condition of interest is r_{max}, only the right side of the inequality need be used. Therefore,

$$\hat{T}\, K'(T')^{-\alpha} = .5$$

solving for K' yields

$$K' = \frac{(.5)(T')^{\alpha}}{\hat{T}}$$

Thus, given the number of different or active thesaurus terms, T', and the total number of term occurrences, $\hat{T}$, one can estimate K' by assuming an α, or estimate α assuming a K'. According to Booth (10), Zipf (153), and Mandelbrot (93-95), $\alpha \approx 1$. Since more is known about the range of α than K and all that is needed is a "quick" approximation, an $\alpha \approx 1$ will be used. With $\alpha = 1$,

$$K' = \frac{(.5)(307)'}{(14)(102)}$$

$$\approx 0.1$$

Note, if f(r) instead of P(r) were being estimated, then

$$K \approx 150.$$

With α and K estimated, the next step is to determine B.

The simplest way to estimate B is at the intercept $f(r=1) = f(i)_{max}$, where B is obviously not negligible because r = 1. Solving

$$f(r) = K(r + B)^{-\alpha}$$

for B, yields,

$$B = \left(\frac{K}{f(r)}\right)^{1/\alpha} - r.$$

For, $K \approx 150$, $f(r) \approx 30$, $r = 1$ and $\alpha \approx 1$, the estimate for B is 4 to 4.5 depending on whether $\alpha = 1$ or 0.9, respectively.

The comparison of MEZ values and the actual term frequency--rank distribution, for the test sample, is shown in Fig. 5.9 and tabulated in Table 5.4.

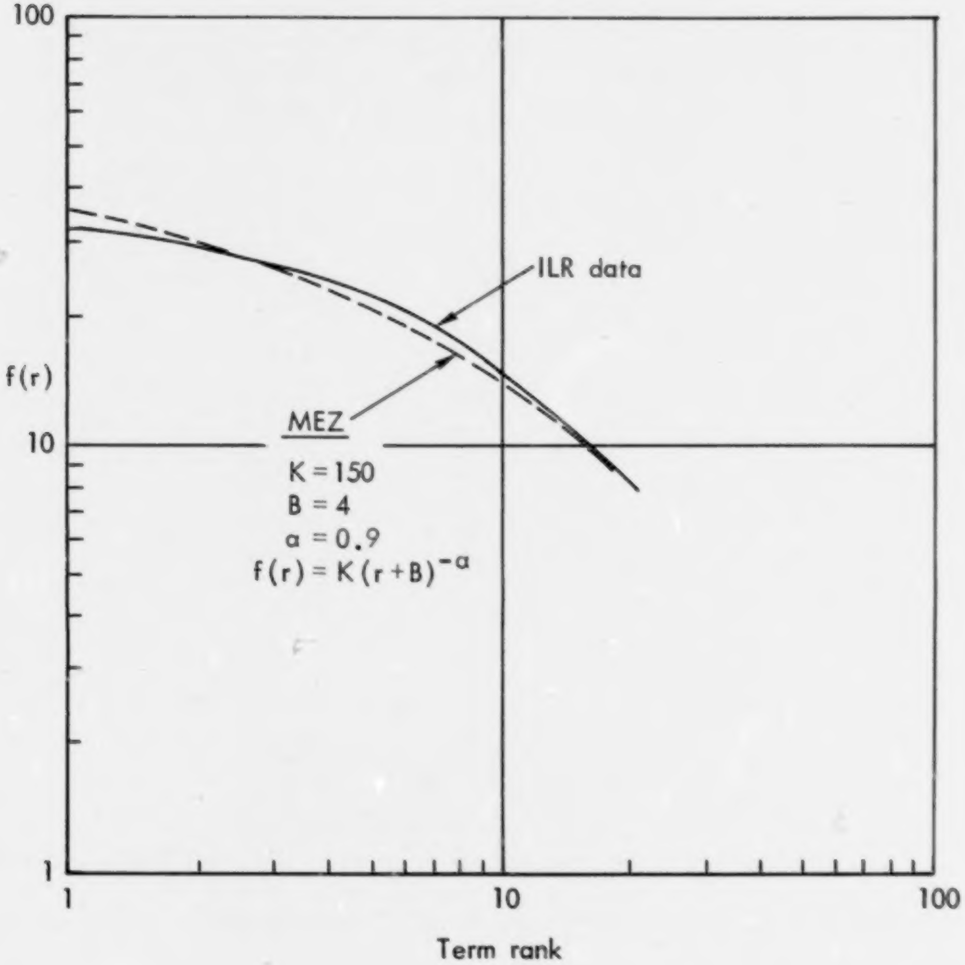


Fig.5.9 — Comparison of ILR test sample term frequency of uses versus rank distribution with the M-E-Z canonical form

Table 5.4

COMPARISON OF MEZ VALUES WITH ACTUAL TERM USAGE VERSUS RANK VALUES FOR THE TEST SAMPLE

Rank	ILR Test Sample	MEZ[a] Values
1	32	35.2
2	27	29.9
3	26	26.0
4	24	23.1
5	22	20.8
6	21	18.9
7	20	17.3
8	17	16.0
9	16	14.9
10	15	14.0
11	14	13.1
12	13	12.4
13	12	11.7
14	11	11.1
15	10	10.6

[a] $K = 150$; $B = 4.5$; $\alpha = 0.9$.

On the basis of this empirical evidence, the hypothesis that the term-usage versus rank relationships are closely approximated by the MEZ canonical form is accepted.

5.3.3 Depth of Indexing Distribution

The depth of indexing distribution is an additional DRS characteristic that can be used to determine statistical similarities between DRSs. The distribution is derived from the DXT distribution (it is the distribution of the row marginals) as indicated in Fig. 5.2.

The indexing density distributions for the test system and the two keyword systems employed by Litofsky (90) are shown in Figs. 5.10 and 5.11 respectively. As for the term usage versus rank distribution, it would be very desirable to represent the depth of indexing distribution by a canonical form. While this exercise is not carried out here, a suggested canonical form is noted in Chapter 7.

5.3.4 The Term-Term Co-occurrence Distribution

The term-term (TXT) matrix is derived from the DXT matrix as shown in Fig. 5.12. For the test system, the TXT matrix is quite sparse (≈82 percent). The non-zero integer entries indicate the number of instances in which the two terms, defining the intersection, are used as common or co-descriptors for documents in the corpus.

Three hypotheses have been put forward regarding the characteristics of the TXT matrix. Each hypothesis will be stated and then analyzed. The first case is:

5.3.4.1 Term Independency. The TXT matrix is not generated by a process which selects terms for assignment independent of one another.

A prevalent assumption in previous analyses is that the descriptor terms in the system thesaurus are assigned independent of one another to documents in the corpus. The often stated qualification is that while this assumption of independency is not exactly satisfied, it is a reasonable approximation. It does not appear that this assumption has ever been statistically tested. Perhaps a complicating factor is that the convenient chi-square test for goodness of fit is not

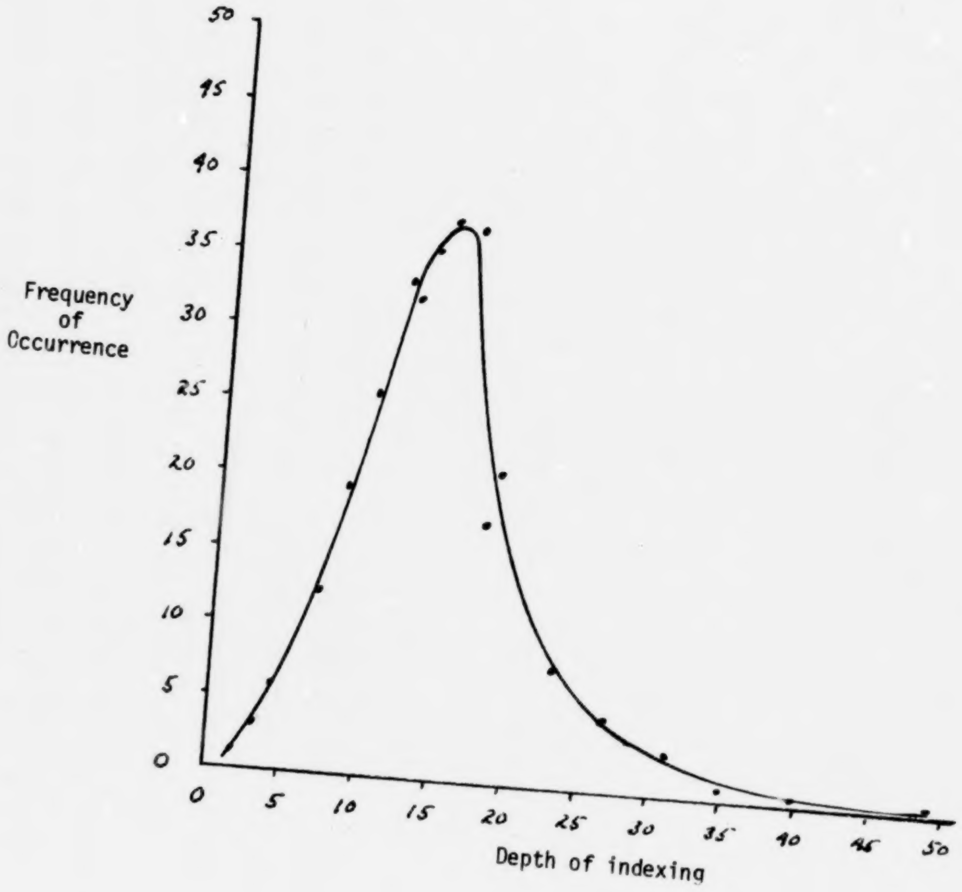


Fig. 5.10 -- Depth of indexing distribution for test sample

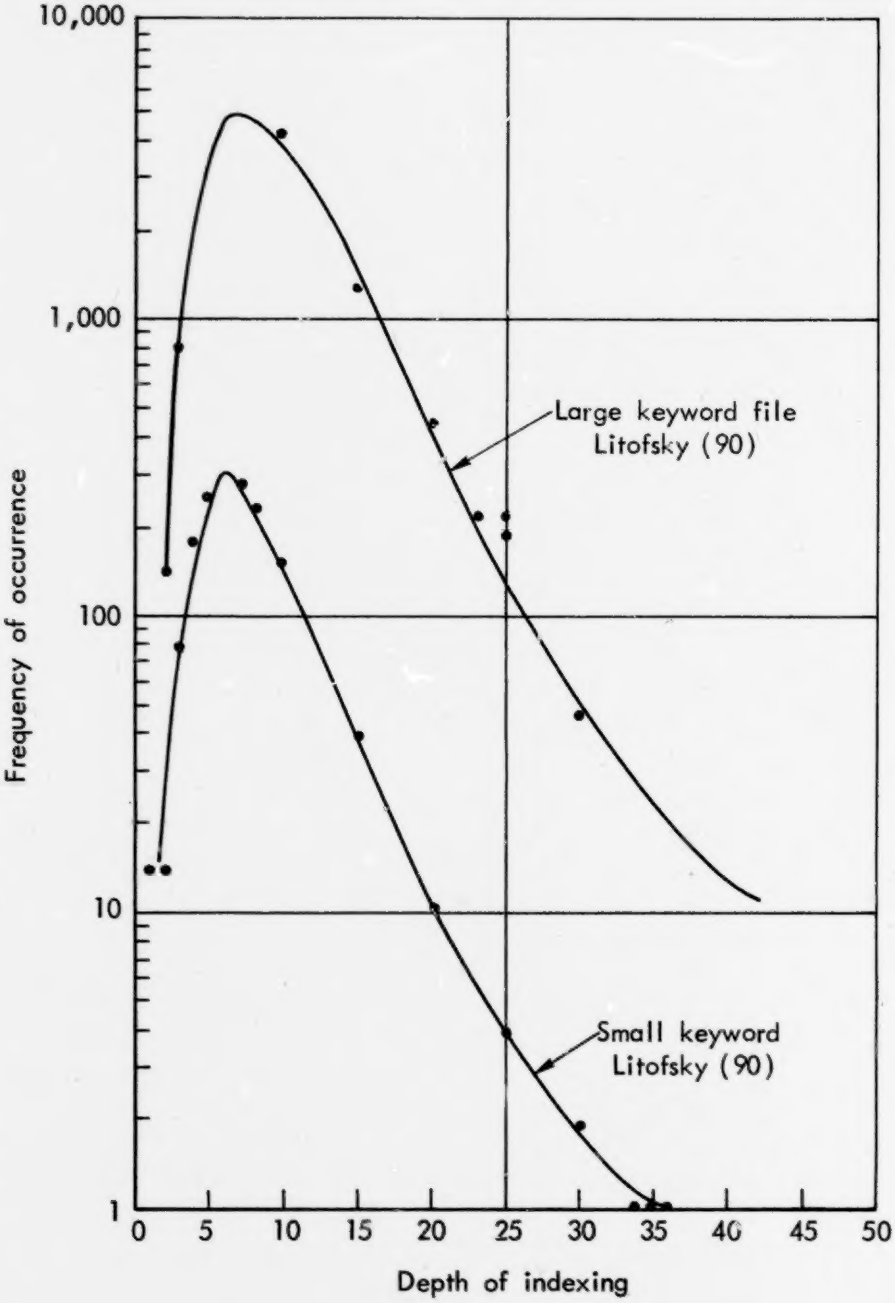
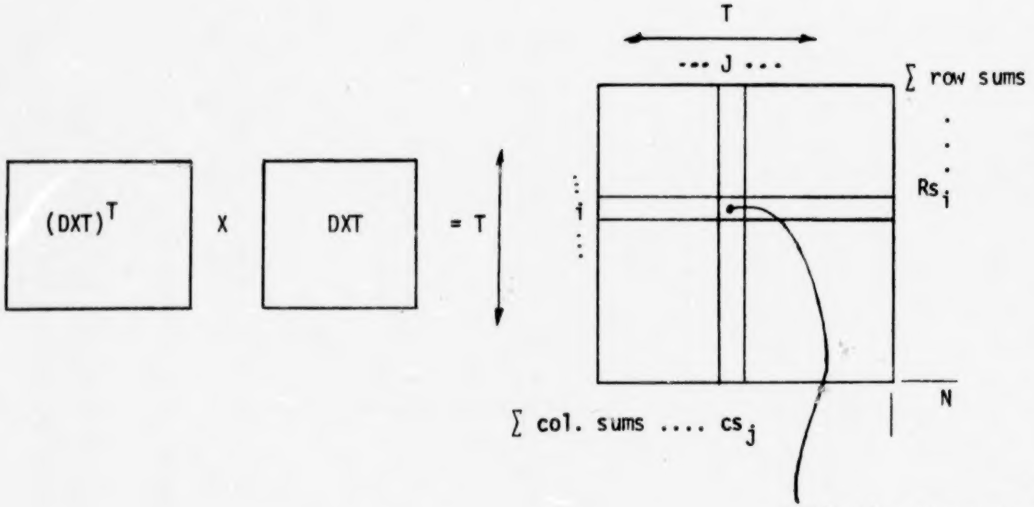


Fig. 5.11 — Depth of indexing distribution for the systems investigated by Litofsky (90)


(DXT)^T
X
DXT
= T
T
... J ...
i
∑ row sums
Rs_i
N
∑ col. sums .... cs_j

$TXT(i,j)$ = the number of documents (in cell i,j)

R_{si} = sum of elements in row i

c_{sj} = sum of elements in column j

$$N = \sum_{i=1}^{T} Rs_i = \sum_{j=1}^{T} cs_j$$

Fig. 5.12 -- Properties of the TXT matrix

appropriate in this case. This is so because the DXT matrix, as defined for the DRSs of interest, is binary and very sparse (i.e., a matrix condition in which the number of elements whose value is zero equals or exceeds the number of elements whose value is near-zero). Thus the theoretical limitation of the chi-square test, which requires that the expected value of the sample of population elements to be tested must be at least equal to 5, is not satisfied. Hence, the chi-square test cannot be used to statistically ascertain whether the DXT matrix is or is not generated as though the descriptors are assigned independent of one another. This situation also holds for the TXT matrix. Even though there are TXT(i,j) which exceed 5, there are many elements that do not because the TXT matrix is also sparse;* this necessarily follows because

$$\mathrm{TXT} = (\mathrm{DXT})^T \cdot (\mathrm{DXT}), \quad \text{and DXT is sparse.}$$

In lieu of the chi-square, the test elected to apply to accept or reject the hypothesis is called the "Generalized-Likelihood-Ratio-Test " (see Mood and Graybill (102)). The Generalized-Likelihood Ratio (GLR) is defined as the quotient

$$\theta = \frac{L(\hat{s})}{L(\hat{o})}$$

where $L(\hat{s})$ = the maximum of the likelihood function in the sample region or space s, with respect to the parameters

$L(\hat{o})$ = the maximum of the likelihood function in the population region or space o, with respect to the parameters and, -2 Log θ is defined as a chi-square variate.

---

*However, it is easily shown that TXT is never more sparse than DXT.

The null hypothesis of interest is that the descriptor terms are assigned independent of one another for each document-descriptor set. When H_0 is true, -2 Log θ is approximately distributed as chi-square with N degrees of freedom when M is large. Thus the null hypothesis can be tested by computing -2 Log θ and comparing it with the desired level of significance of chi square. If -2 Log θ exceeds the chi-square level, H_0 will be rejected, otherwise H_0 will be accepted.

Given the DXT matrix, as illustrated in Fig. 5.13, the desire is to show that the assignment of any one of the terms in the matrix is independent of the occurrence of any other term; that is to say, the probability of occurrence of term i is independent of term j. The null hypothesis is:

$$H_0 \; : \; f(n_1,\dots,n_N/H_0) = \prod_{i=1}^{N} P_i^{n_i} q_i^{M-n_i}$$

where P_i = probability of term i occurring n_i times

$q_i = (1-P_i)$

M = the number of documents to be indexed

N = the number of terms in the thesaurus

To test H_0, the GLR θ is computed, where

$$\theta = \frac{L(\hat{s})}{L(\hat{o})}$$

and,

$$L(\hat{s}) = \sup_{\tilde{p} \epsilon H_0} \sum_{i=1}^{N} \left(\frac{n_i}{M}\right)^{n_i} \left(\frac{M-n_i}{M}\right)^{M-n_i}$$

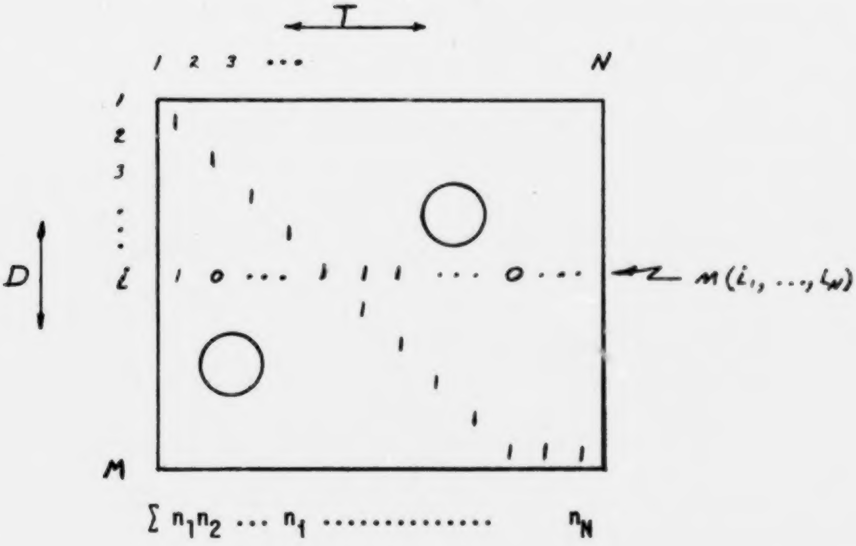

T
1 2 3 ...
N
1
2
3
D
i
M
$m(i_1, \ldots, i_N)$
$\sum n_1 n_2 \ldots n_i \ldots\ldots n_N$

Fig. 5.13 -- Illustration of a sparse DXT matrix with statistical parameters employed in the GLR test

where in the context of this problem, n_i/M are normalized frequencies and are taken to be sufficiently representative of the probabilities of occurrence of the descriptor terms. Also,

$$\tilde{P} = \text{Vector of } (P_1, \ldots, P_n)$$

which maximizes the function $L(\hat{s})$. In this case, the empirically observed frequencies of occurrences or the "best estimates" of the elements of $\hat{P}$.

Introducing Logs for ease of computation yields

$$\text{Log } L(\hat{s}) = \underset{\tilde{P}\varepsilon H_o}{\text{Sup}} \sum_{i=1}^{n} n_i \text{ Log } \frac{n_i}{M} + (M-n_i) \text{ Log}\left(\frac{m-n_i}{M}\right)$$

where the normalized frequencies, f_i*

$$f_i = \frac{n_i}{M}$$

can be substituted, giving

$$\text{Log } L\ (\hat{s}) = \underset{\tilde{P}\varepsilon H_o}{\text{Sup}} \sum_{i=1}^{N} n_i \text{ Log } f_i + (M-n_i) \text{ Log}(1-f_i)$$

Now it is necessary to compute, $L(\hat{o})$

$$L(\hat{o}) = \underset{P_{i_1},\ldots,i_{11}}{\text{Sup}} f(n_1, \ldots, n_N)$$

$$L(\hat{o}) = \underset{P_{i_1},\ldots,i_N}{\text{Sup}} \sum_{i_1,\ldots,i_N}^{N} (P_{i_1},\ldots,i_N)^{n(i_1,\ldots,i_N)}$$

---

*The implicit assumption is that a term can be assigned only once to a document. Therefore, the maximum frequency of use of any term is the number of documents in the corpus, M.

where $P_{i_1,\ldots,i_N}$ is the probability that a randomly chosen document has descriptor vector $n(i_1,\ldots,i_N)$ and is defined by

$$P_{i_j} = \left(\frac{n(i_1,\ldots,i_N)}{M}\right)$$

Substituting, and introducing the Log for convenience yields,

$$\text{Log}\ (\hat{o}) = \overset{\text{Sup}}{P_{i_1,\ldots,i_N}}\, \mathfrak{f} = \sum_{i_1,\ldots,i_N}^{N} n(i_1,\ldots,i_N)\ \text{Log}\left(\frac{n(i_1,\ldots,i_N)}{M}\right)$$

Assuming, that the identical occurrence of $n(i_1,\ldots,i_N)$ for more than a few documents is not a very likely event,[*] then Log $L(\hat{o})$ can be simplified as follows

$$\text{Log}\ L(\hat{o}) = \sum_{j=1}^{K} jd(j)\ \text{Log}\left(\frac{j}{M}\right); \quad \text{for} \quad j \leq M$$

where K is the maximum number of congruent document vectors, and d(j) is the number of descriptor vectors which correspond to exactly j documents. In fact, the usual case (of which the test system is an example), K=1, and the above relationship reduces to

$$\text{Log}\ L(\hat{o}) = M\ \text{Log}\ \frac{1}{M}$$

Therefore, the expression to be evaluated is:

---

[*] The most unlikely event is when the identical occurrences of $n(i_1,\ldots,i_N)$ is M, which means the corpus consists of M "identical" items -- in so far as the thesaurus subject delineation of concepts/subjects is concerned.

$$\text{Log } \theta = \sum_{i=1}^{N} n_i \text{ Log } f_i + (M-n_i) \text{ Log}(1-f_i) - M \text{ Log } \frac{1}{M}$$

and, -2 Log θ is the chi square variate of interest with N degrees of freedom.* Since, for this analysis, $N = 370$, the normal approximation to the chi square distribution is used.

For the test sample, $-2 \text{ Log } \theta \simeq 850$ which is larger than the normal approximation to the chi square, which at the .005 level, $X^2 \simeq 480$. Therefore, the hypothesis of term independency is rejected.

5.3.4.2 Term-Term Co-occurrence Factor. The next hypothesis to test is whether the co-occurrence of two terms is directly proportional to a function of the frequencies of use of the terms.

In Chapter 4, two candidate functions were proposed:

I. $$TXT(i,j) = \gamma\left(\frac{f(i)\cdot f(j)}{D}\right)$$

II. $$TXT(i,j) = \gamma_1 \frac{RS(i)\cdot CS(j)}{\Sigma RS}$$

where $f(i)$ = the frequency of use of term i

$TXT(i,j)$ = the value of the intersection of term i and j

$RS(i)$ = the sum of the entries in row i

$CS(i)$ = the sum of the entries in column j

D = the number of documents indexed.

The relationships of the above functions and variables and the TXT matrix are shown in Fig. 5.12.

The variables of interest in the above equations are the γ's. That is, in order for the estimations to be useful, the distribution of values for γ must be stable and stationary. Therefore the forms of the relationships that will be analyzed are:

---

*The variable P_{i_j} is allowed to vary over the range 0 to 1, with $j=1,\ldots,N$.

$$\text{I}' \qquad \gamma = \frac{TXT(i,j)}{\frac{f(i)\cdot f(j)}{D}} = \frac{\text{Actual}}{\text{Theoretical}}$$

$$\text{II}' \qquad \gamma_1 = \frac{TXT(i,j)}{\frac{RS(i)\cdot CS(j)}{\Sigma RS(i)}} = \frac{\text{Actual}}{\text{Theoretical}}$$

A computer program was written to analyze a sample of the test DRS TXD distribution. The program generated the TXT(i,j) for every non-zero cell in TXT, computed the values of the candidate function. and the ratio of the actual to theoretical values for γ and γ_1. A small sample is presented in Table 5.5. It is clearly evident that relationship I or I' is superior to relationship II or II'. Function II is very unstable (it has a large variance) and it is not suitable as an estimator of the value of TXT(i,j).

On the other hand, function I is very stable. The plot of theoretical γ versus f(i) in log-log space is always linear, and all the theoretical values of γ for any f(i) can be determined from a knowledge of the relationship of f(1) and the γ's for f(1). An illustration of this relationship is given in Fig. 5.14.

The empirical values of γ for terms with f(i) = 1 to f(i) = 32,* are plotted in Figs. 5.15 to 5.30. As shown, each occurrence or value of γ either falls on the theoretical lower bound or lies above it on

---

*The highest term frequency of use in the data sample.

Table 5.5

COMPARISON OF γ AND γ_1 FACTORS

f(i) x f(j)	1 × 1		1 × 2		1 × 3		1 × 4		3 × 3		3 × 4	
Sample Point	γ	γ_1	γ	γ_1	γ	γ_1	γ	γ_1	γ	γ_1	γ	γ_1
1	102	10.67	51	10.06	34	5.87	25.5	12.50	11.33	7.77	8.50	63.40
2	102	11.01	51	16.77	34	27.09	25.5	16.01	11.33	6.22	8.50	5.40
3	102	11.74	51	17.61	34	27.63	25.5	4.24	11.33	9.56	8.50	2.55
4	102	16.77	51	10.06	34	27.09	25.5	4.19	11.33	12.43	8.50	9.55
5	102	18.54	51	16.77	34	11.84	25.5	6.18	11.33	19.13	8.50	4.02
6	102	20.75	51	17.61	34	49.73	25.5	12.58	11.33	16.58	8.50	5.46
7	102	22.01	51	10.98	34	41.44	25.5	152.77	11.33	49.73	8.50	5.88
8	102	23.48	51	--	34	6.40	25.5	13.72	11.33	9.39	8.50	5.03
9	102	70.44	51	--	34	29.56	25.5	10.18	11.33	16.91	8.50	13.89
10	102	10.67	51	--	34	--	25.5	4.63	11.33	31.70	17.00	7.29

	3 × 5		3 × 6		4 × 4		4 × 5		4 × 7	
	γ	γ_1	γ	γ_1	γ	γ_1	γ	γ_1	γ	γ_1
1	6.80	3.50	5.67	2.46	12.75	3.64	5.10	7.68	3.64	1.29
2	6.80	27.63	5.67	5.67	6.38	4.63	5.10	4.45	3.64	1.70
3	6.80	4.23	5.67	2.99	6.38	2.68	5.10	19.51	7.29	5.36
4	6.80	15.85	5.67	5.98	6.38	3.18	5.10	23.05	3.64	6.94
5	6.80	9.39	11.33	5.23	6.38	5.46	5.10	2.12	7.29	14.55
6	6.80	6.50	5.67	5.46	6.38	7.27	5.10	3.25	3.64	10.91
7	6.80	9.75	5.67	9.66	6.38	10.18	5.10	6.37	3.64	7.28
8	13.60	26.69	5.67	--	6.38	2.65	5.10	16.97	3.64	1.68
9	6.80	11.03	5.67	--	12.75	6.29	5.10	2.10	3.64	2.65
10	13.60	84.53	5.67	--	6.38	4.31	10.20	12.58	7.29	14.38

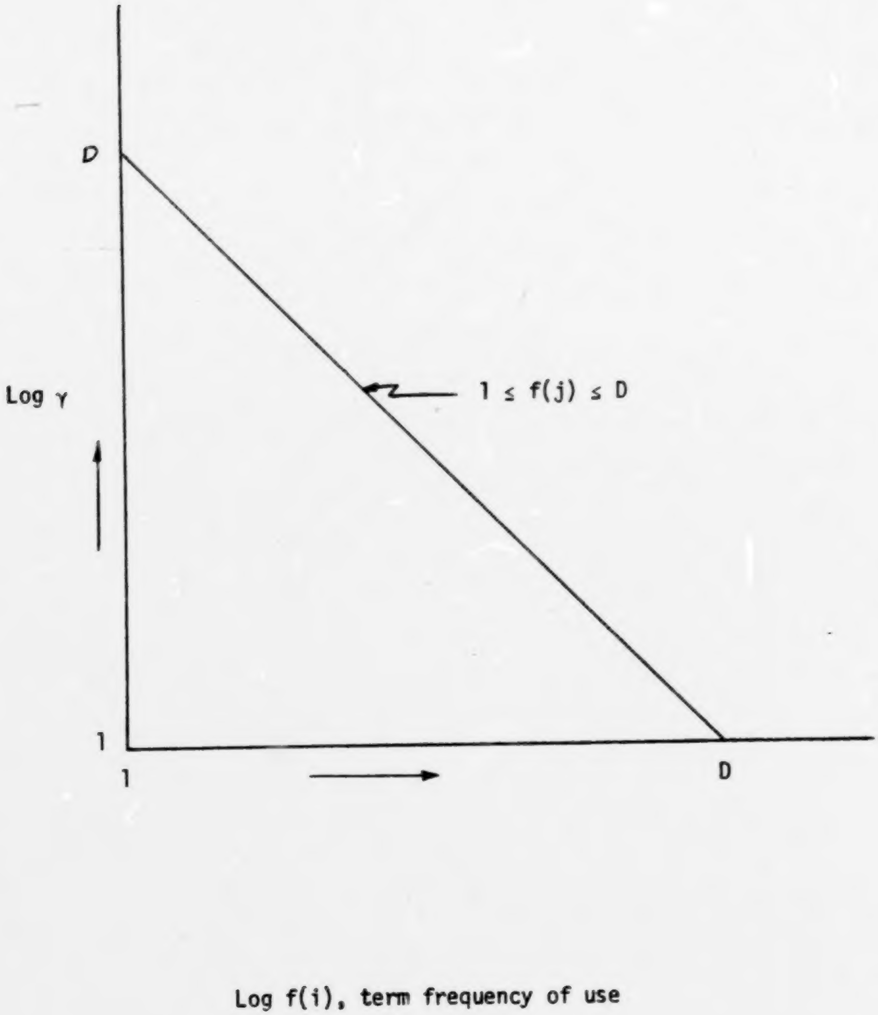


Fig. 5.14 -- Theoretical TXT Prediction Factor-γ

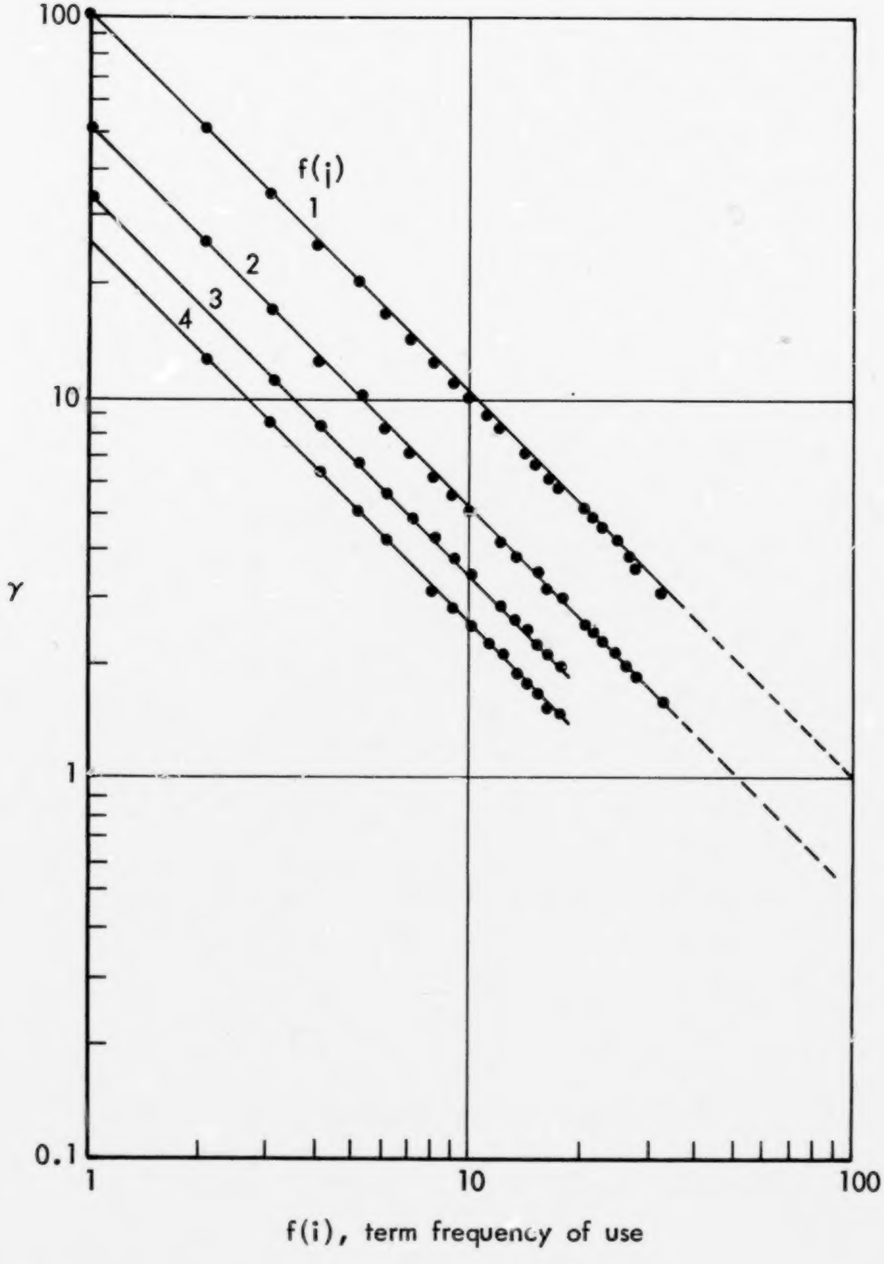


Fig.5.15 — γ factors for $f(j) = 1, 2, 3, 4$ and $1 \le f(i) \le 32$

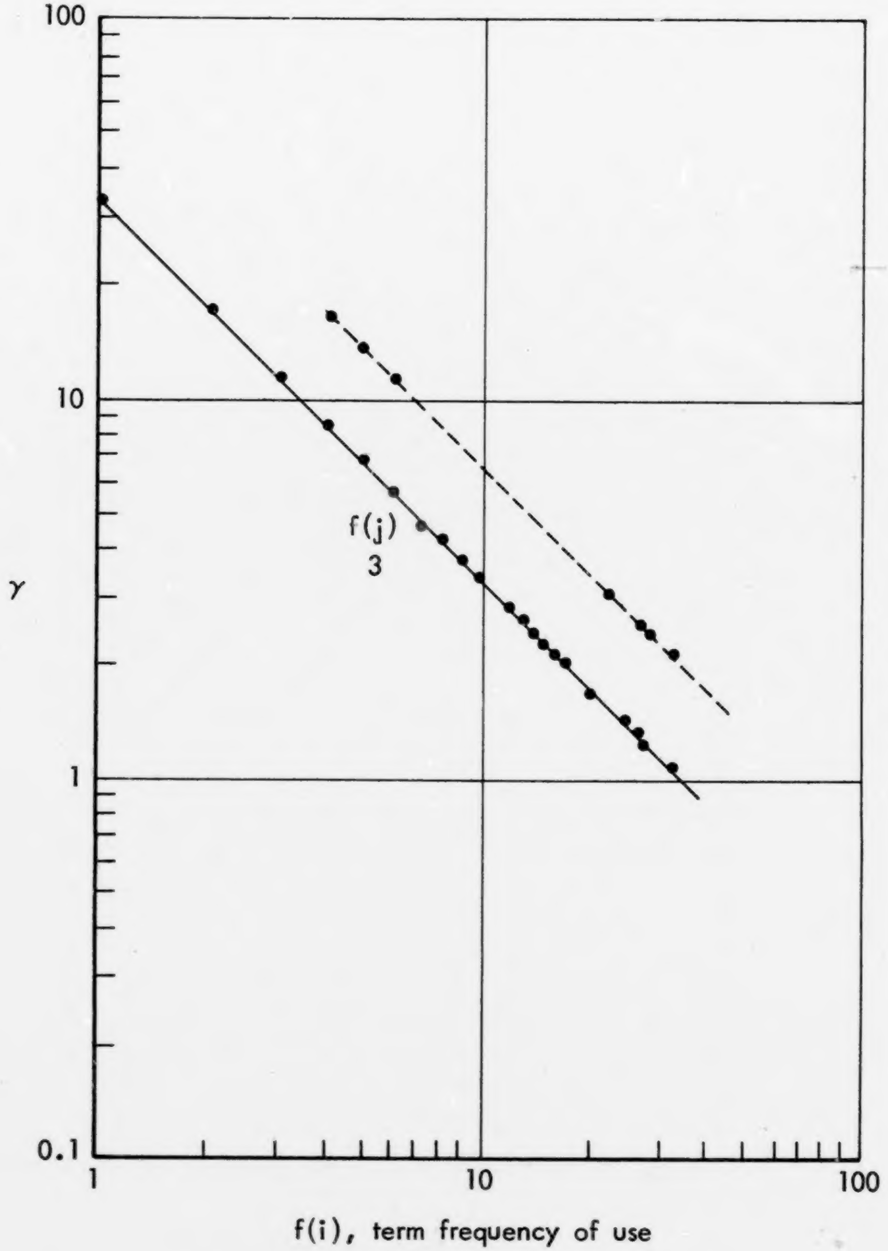


Fig.5.16 — γ factors for f(j) = 3 and 1 ≤ f(i) ≤ 32

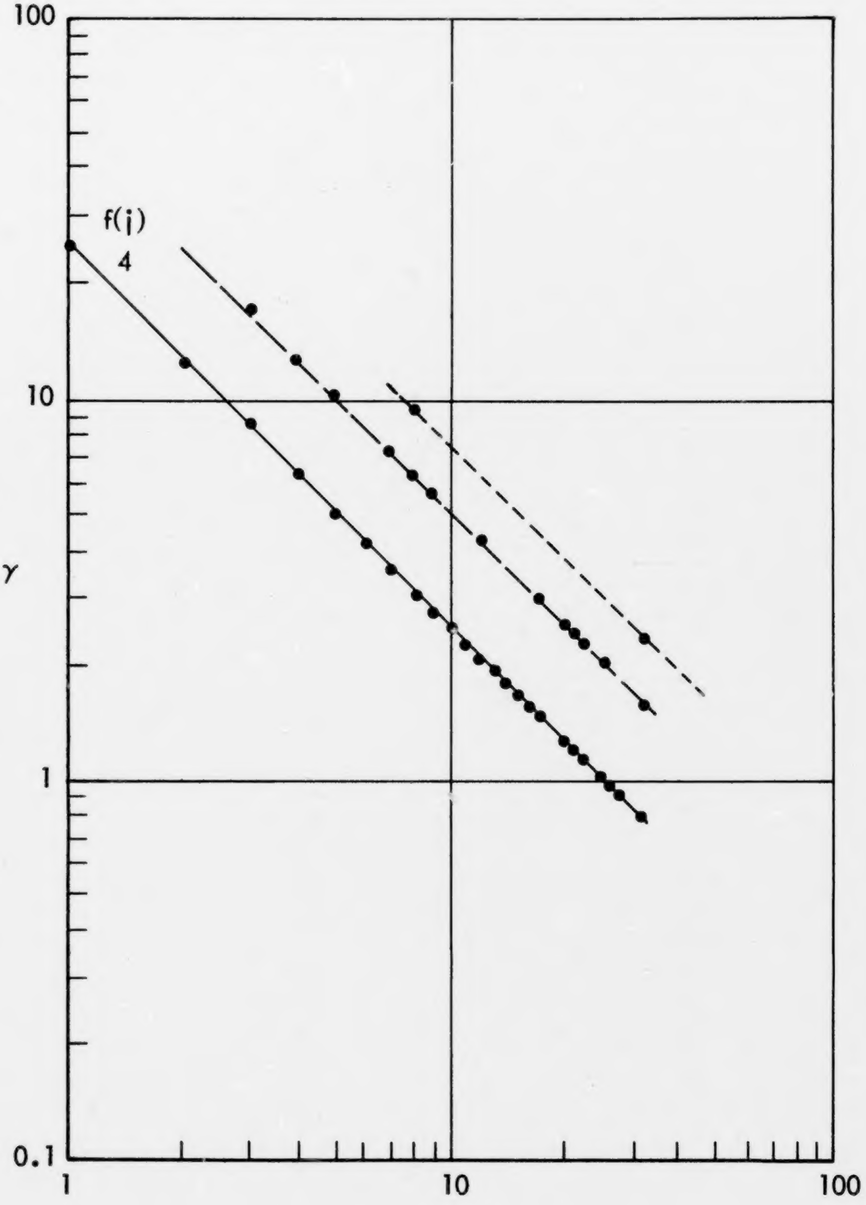


Fig.5.17 — γ factors for f(j) = 4 and 1 ≥ f(i) ≥ 32

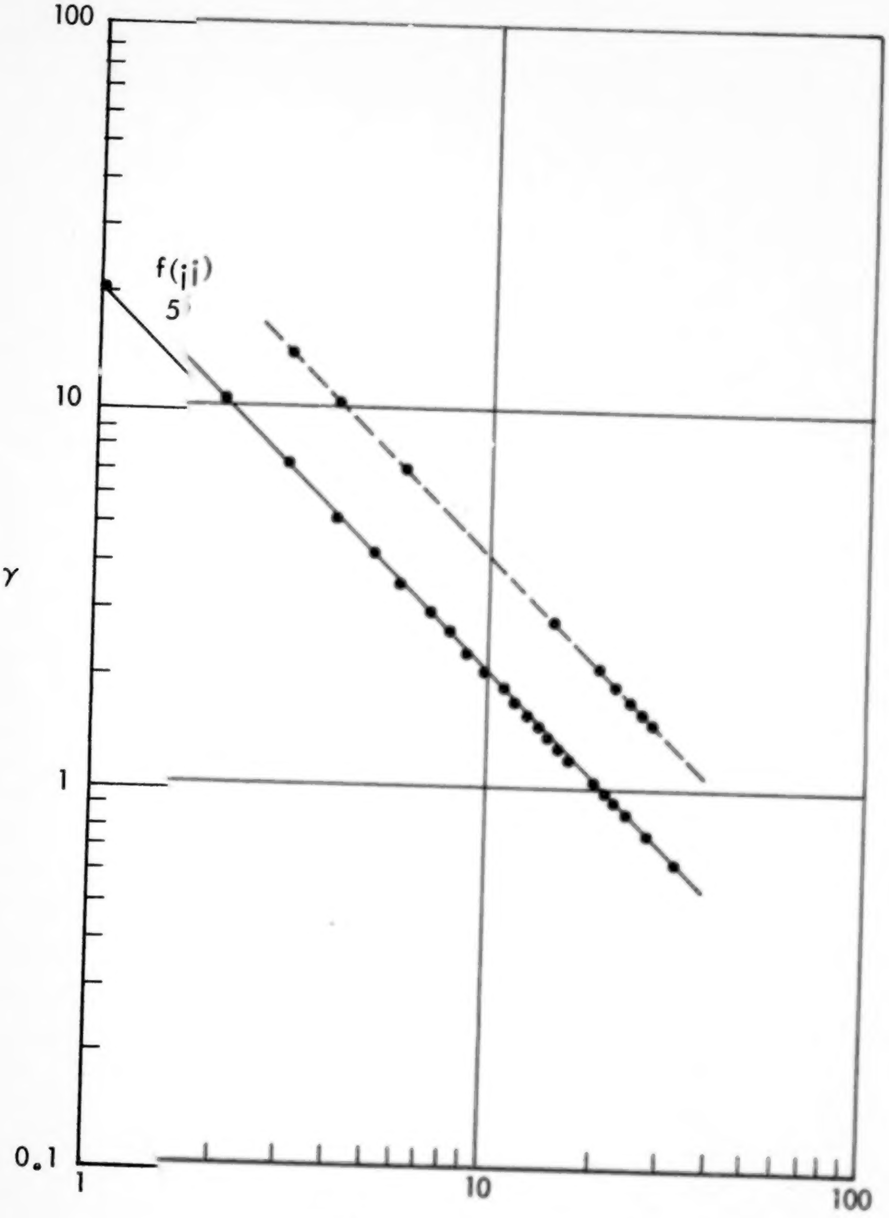


Fig.5.18 — γ factors for f(j) = 5 and 1 ≤ f(i) ≤ 32

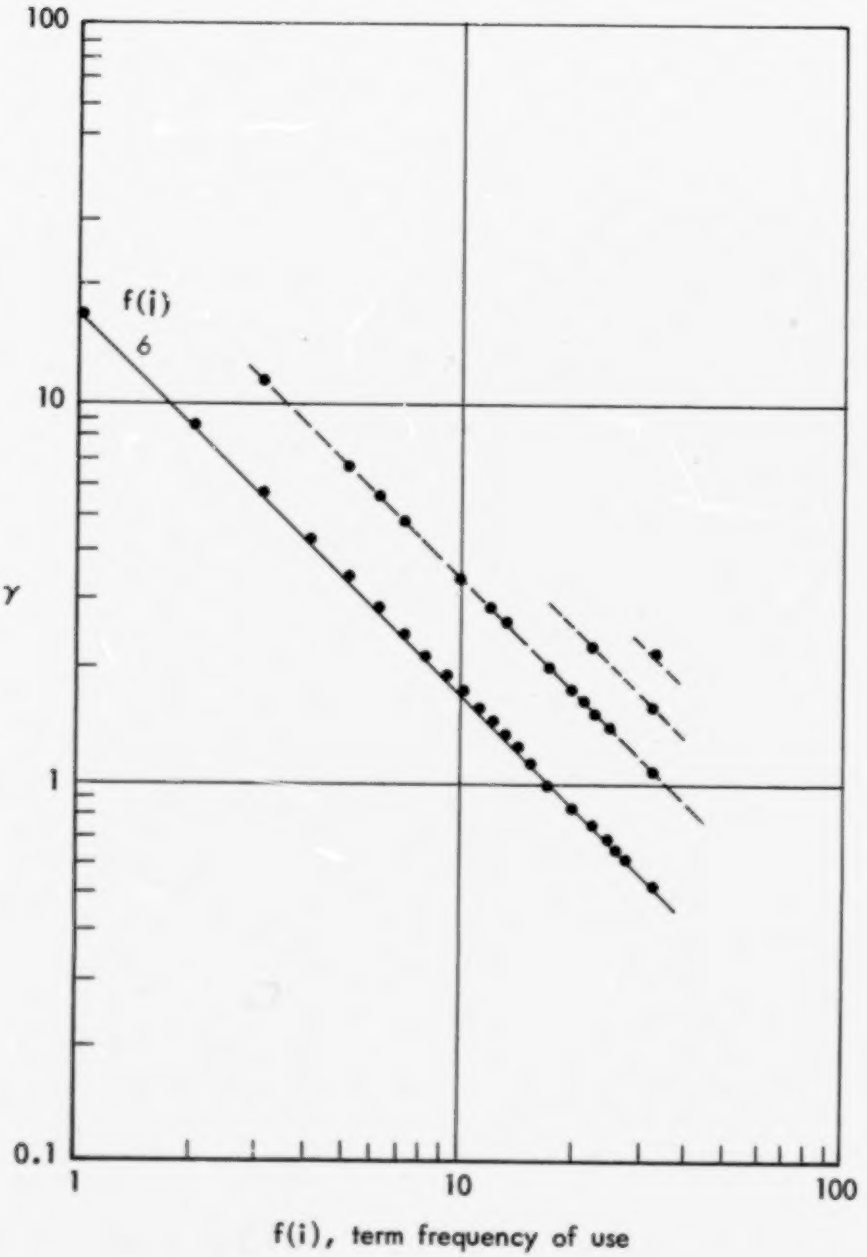


Fig.5.19 — γ factors for f(j) = 6 and 1 ≤ f(i) ≤ 32

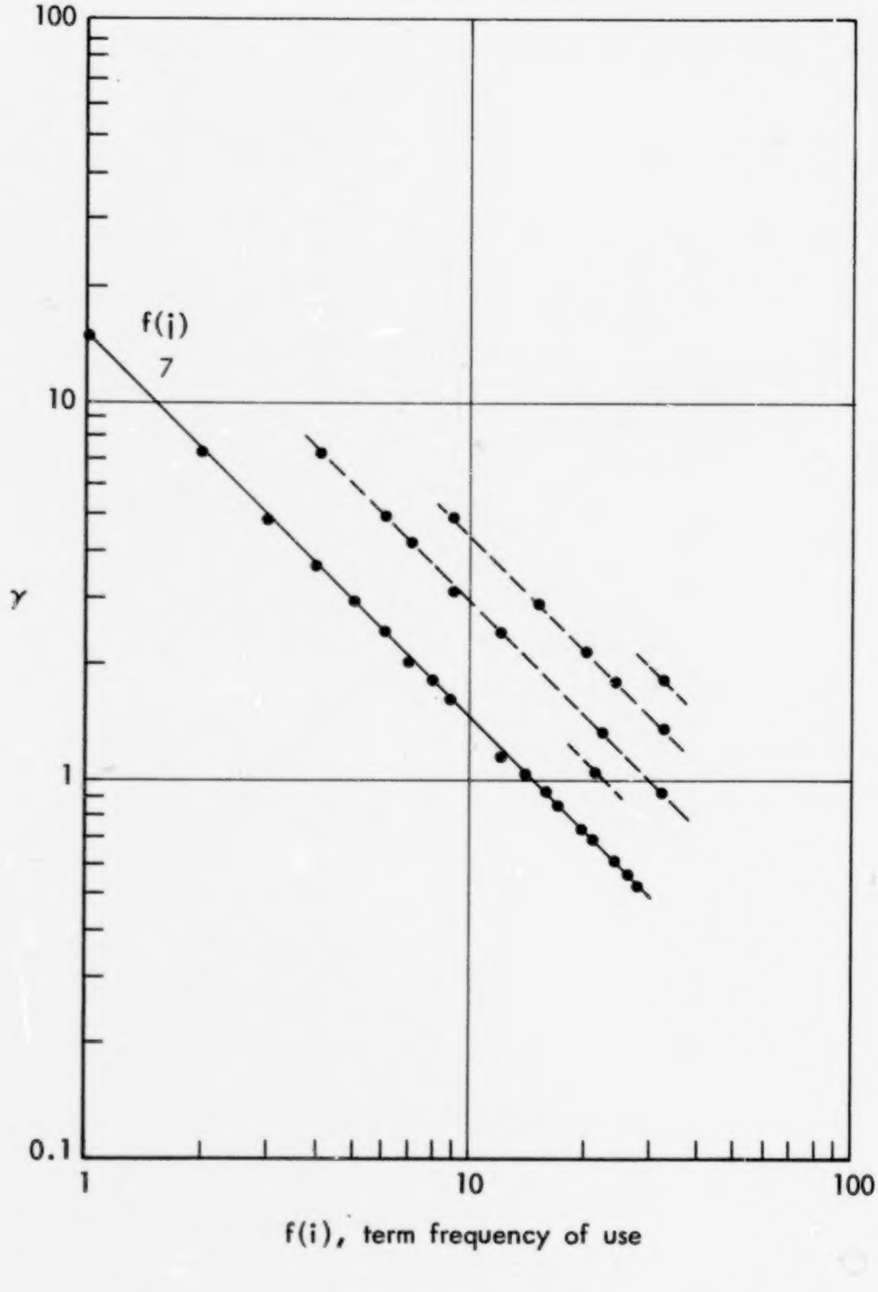


Fig.5.20 — γ factors for $f(j) = 7$ and $1 \leq f(i) \leq 32$

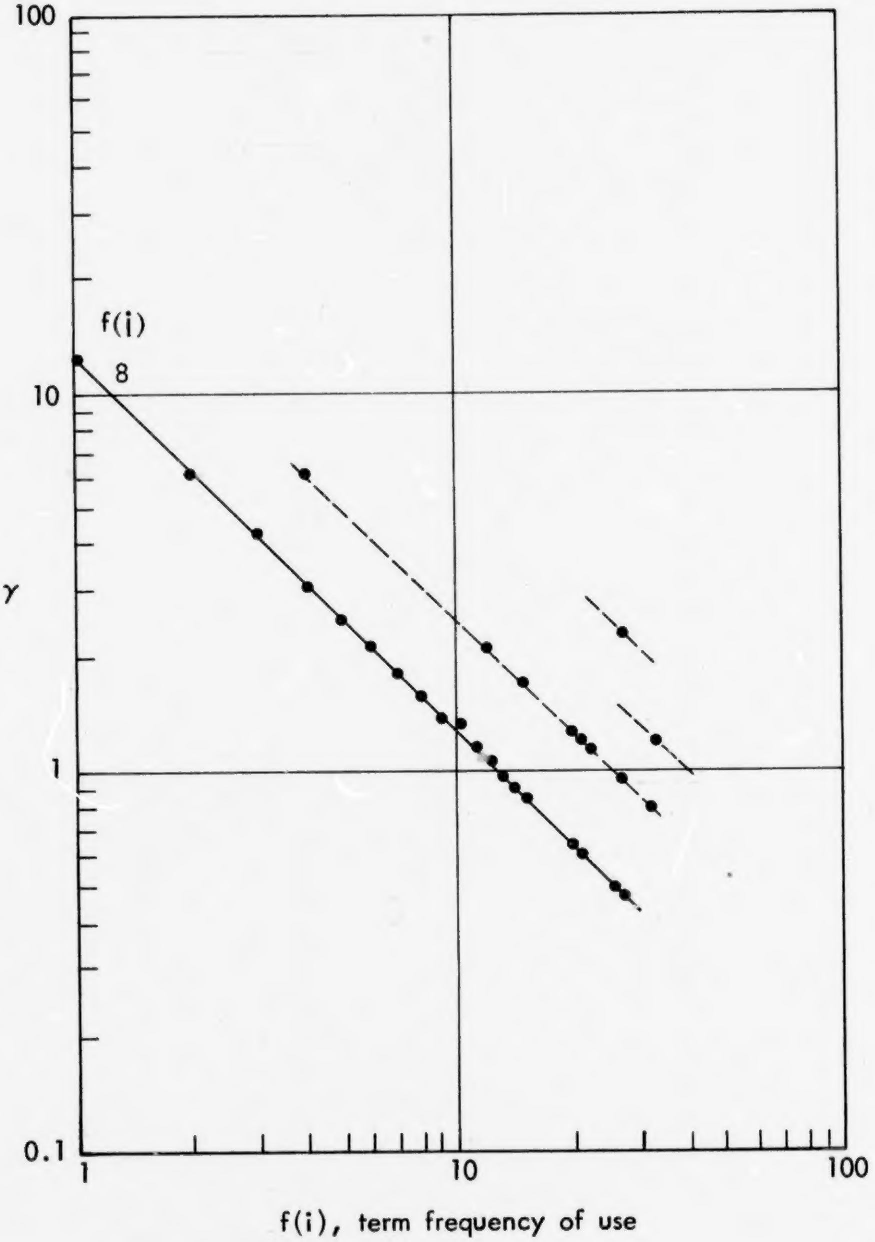


Fig.5.21 — γ factors for f(j) = 8 and 1 ≤ f(i) ≤ 32

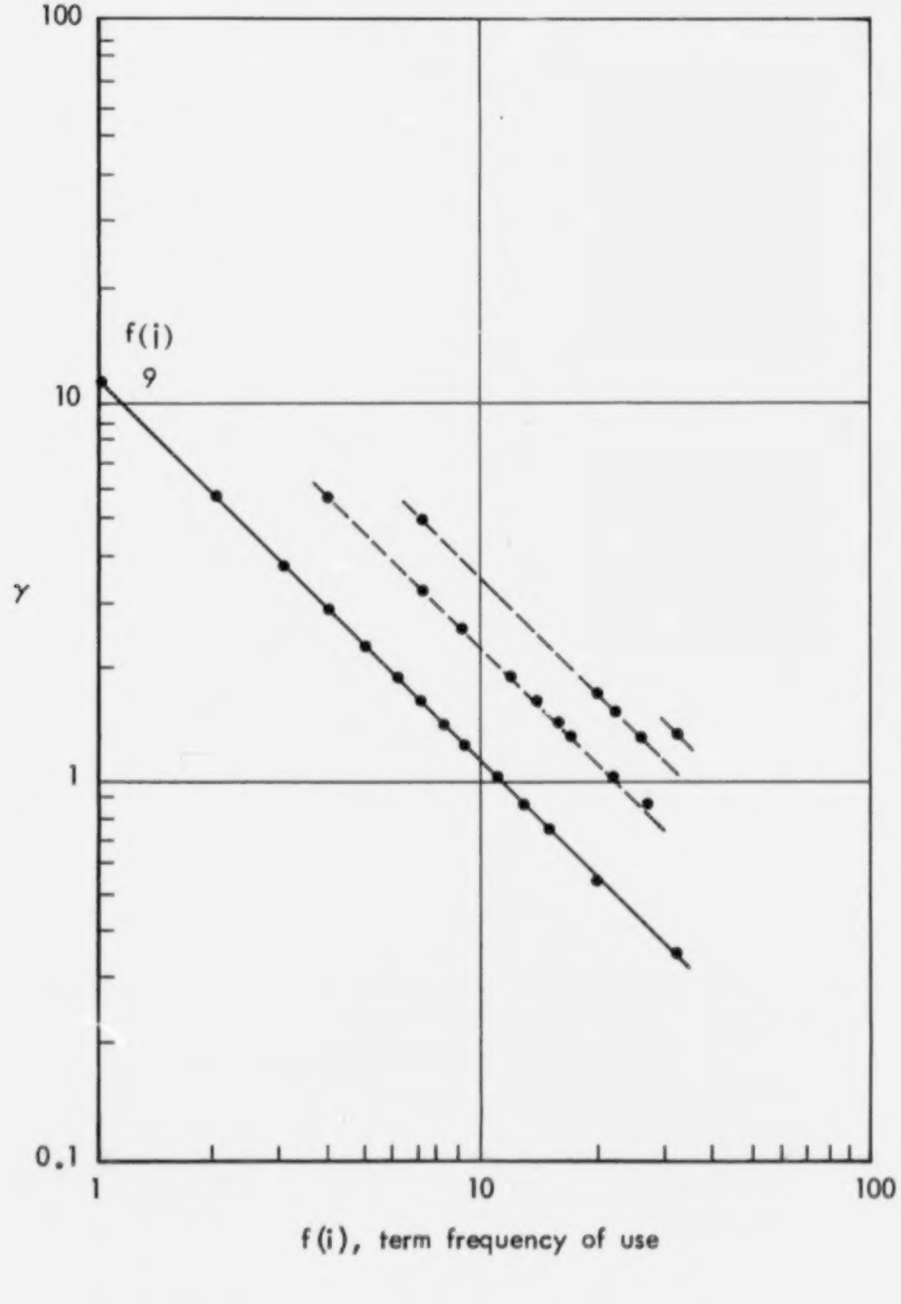


Fig.5.22 — γ factors for $f(j) = 9$ and $1 \leq f(i) \leq 32$

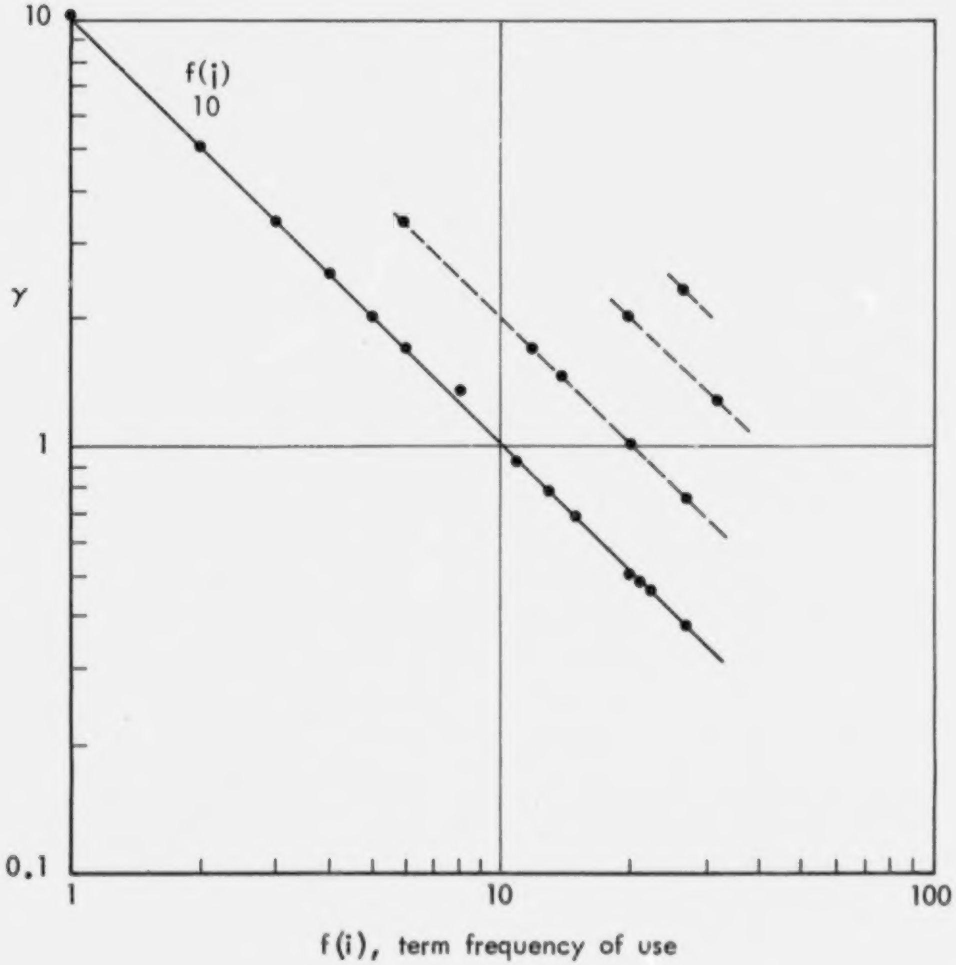


Fig.5.23 — γ factors for $f(j) = 10$ and $1 \leq f(i) \leq 32$

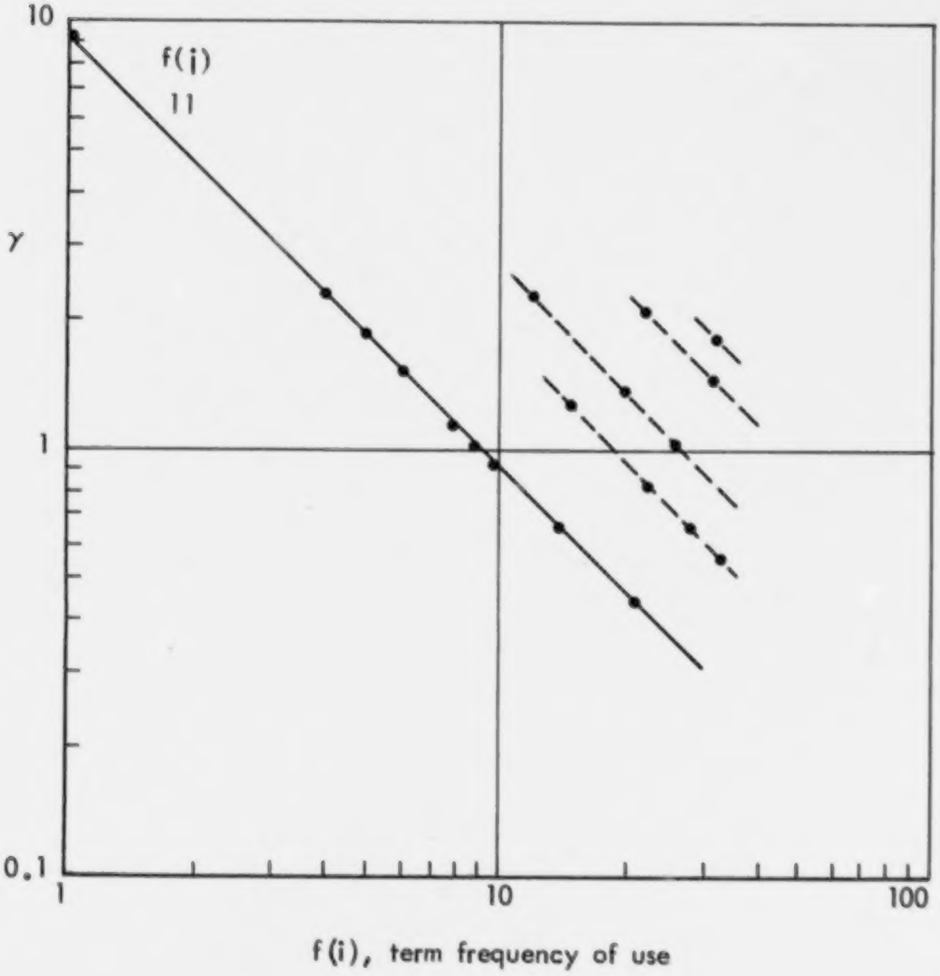


Fig.5.24 — γ factors for f(j) = 11 and 1 ≤ f(i) ≤ 32

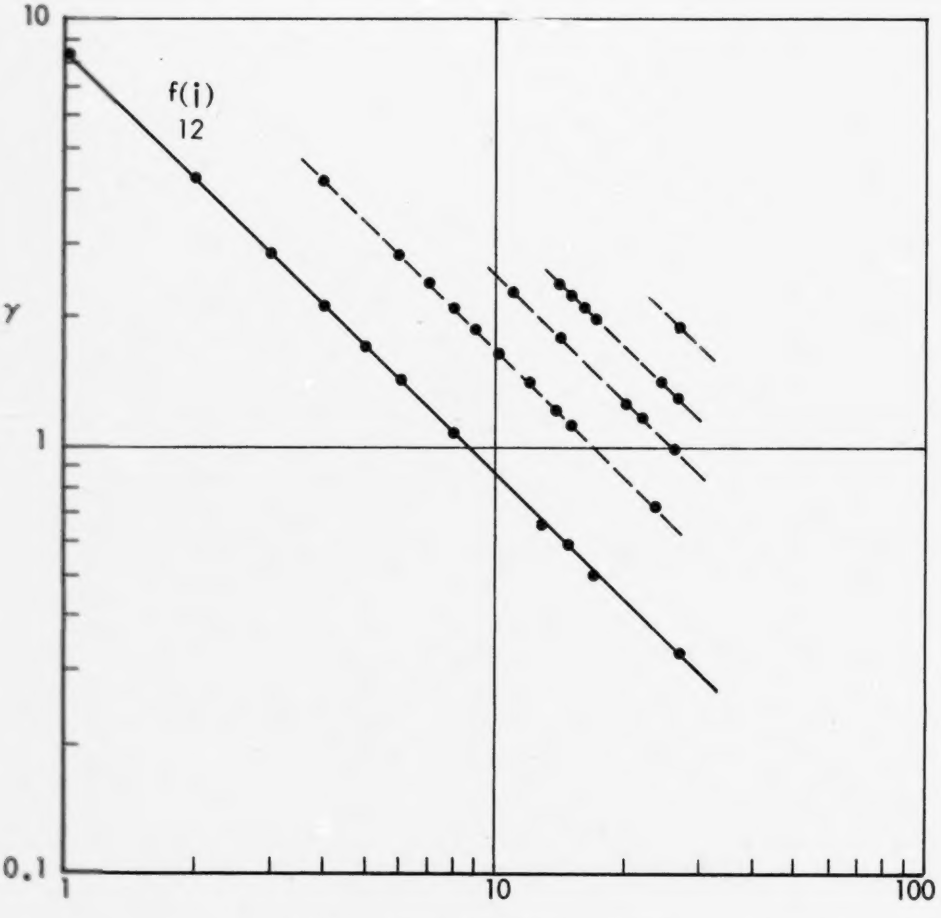


Fig.5.25 — γ factors for f(j) = 12 and 1 ≤ f(i) ≤ 32

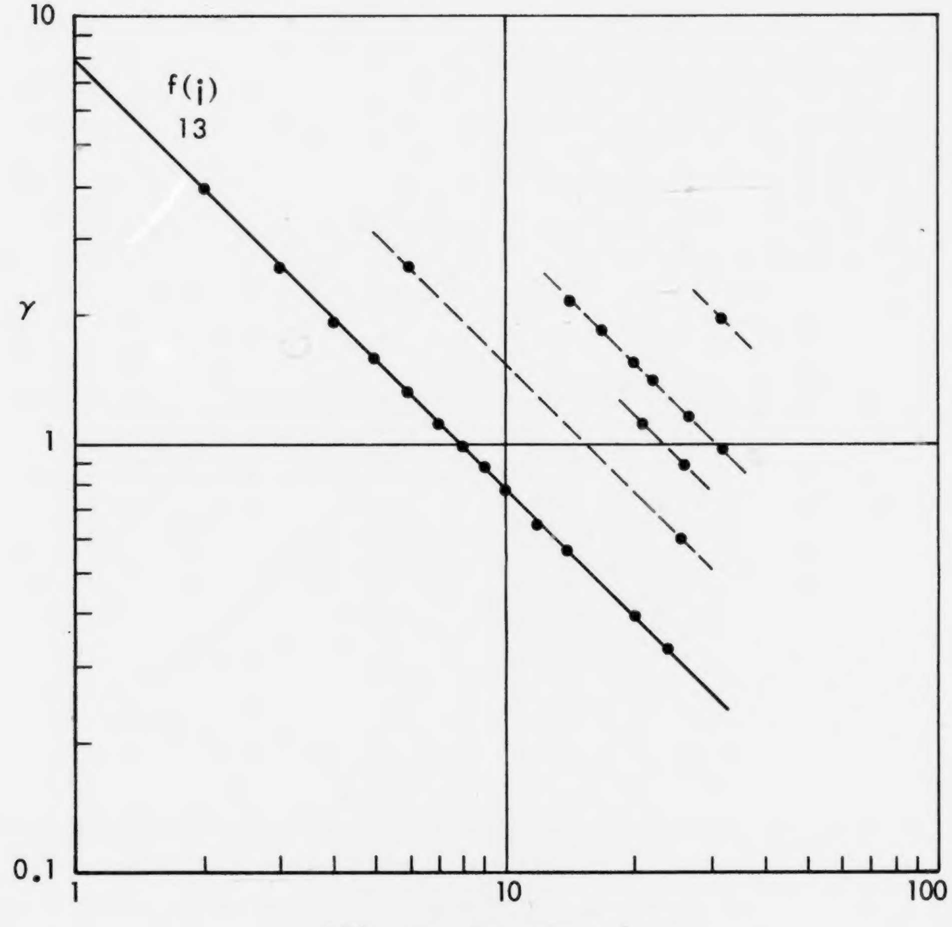


Fig.5.26 — γ factors for $f(j) = 13$ and $1 \leq f(i) \leq 32$

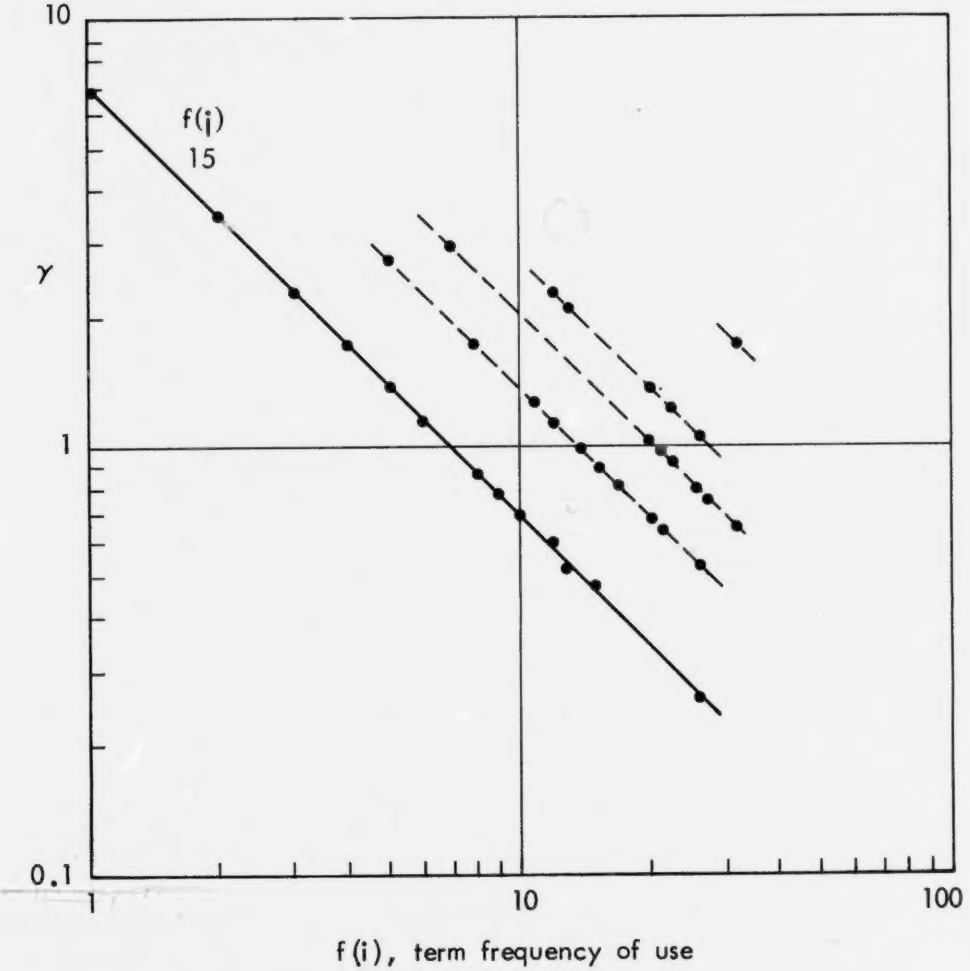


Fig.5.27 — γ factors for f(j) = 15 and 1 ≤ f(i) ≤ 32

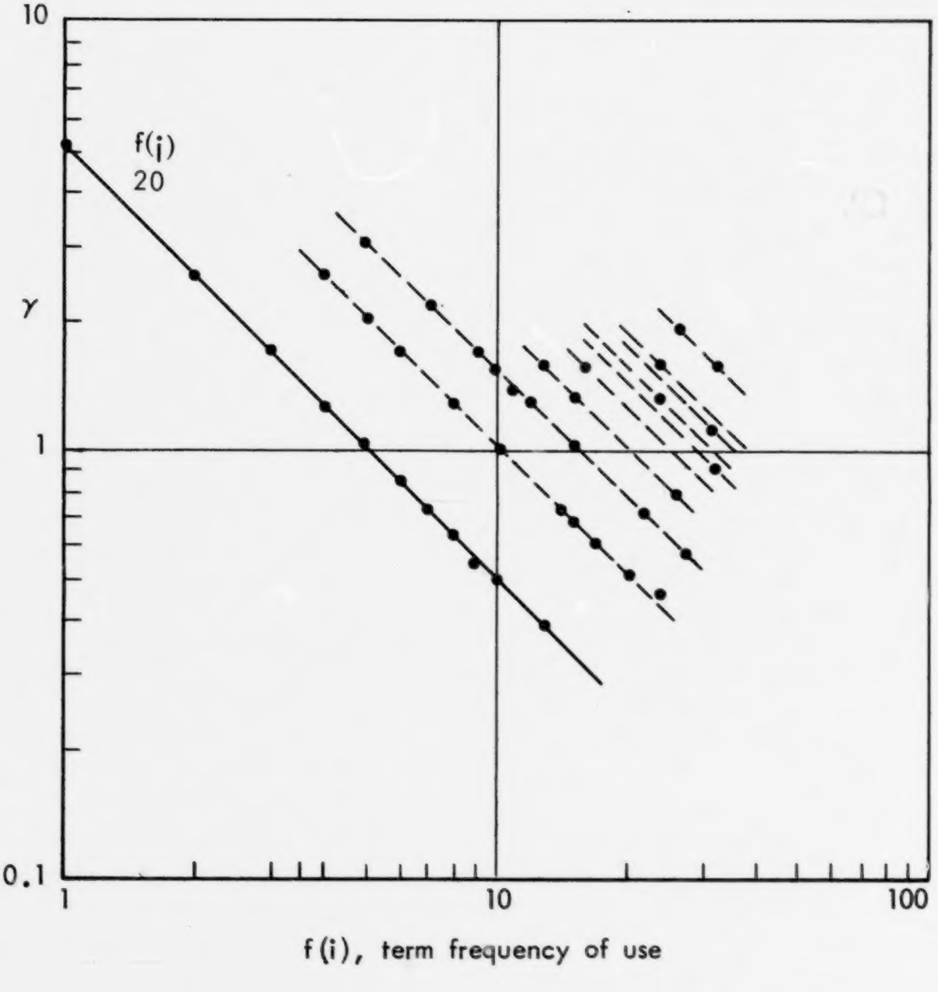


Fig.5.28 — γ factors for f(j) = 20 and 1 ≤ f(i) ≤ 32

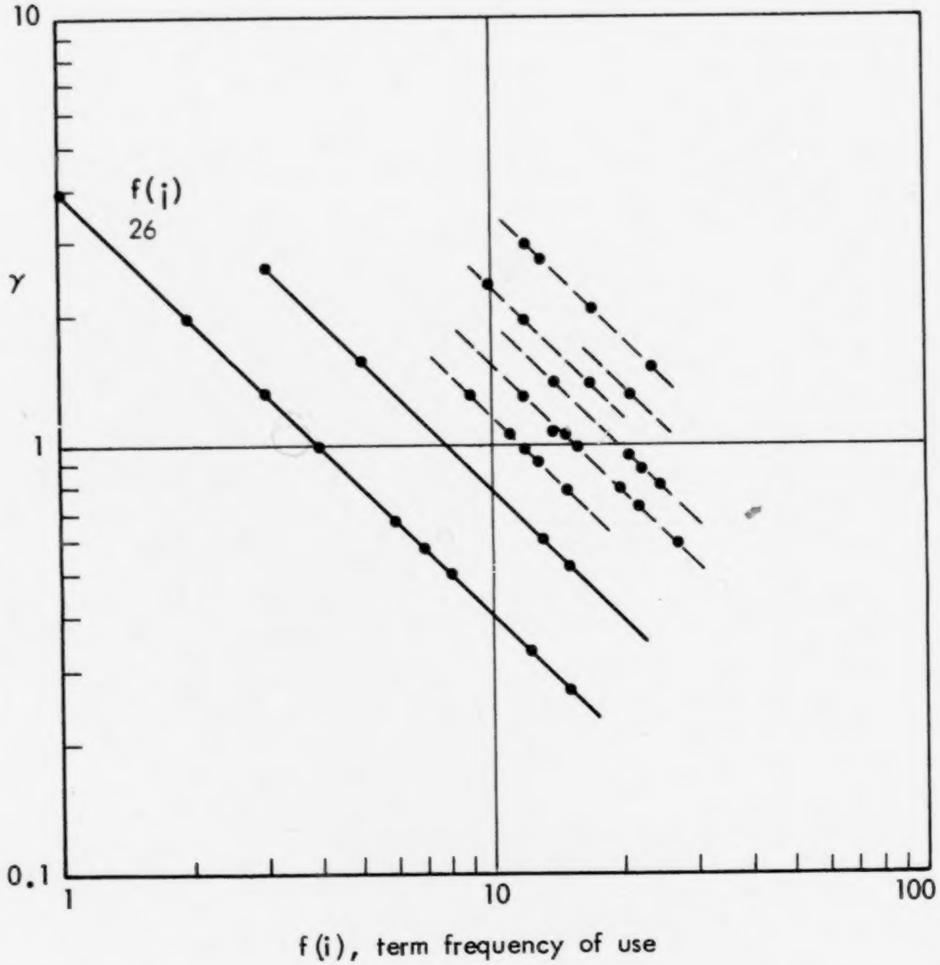


Fig.5.29 — γ factors for f(j) = 26 and 1 ≤ f(i) ≤ 32

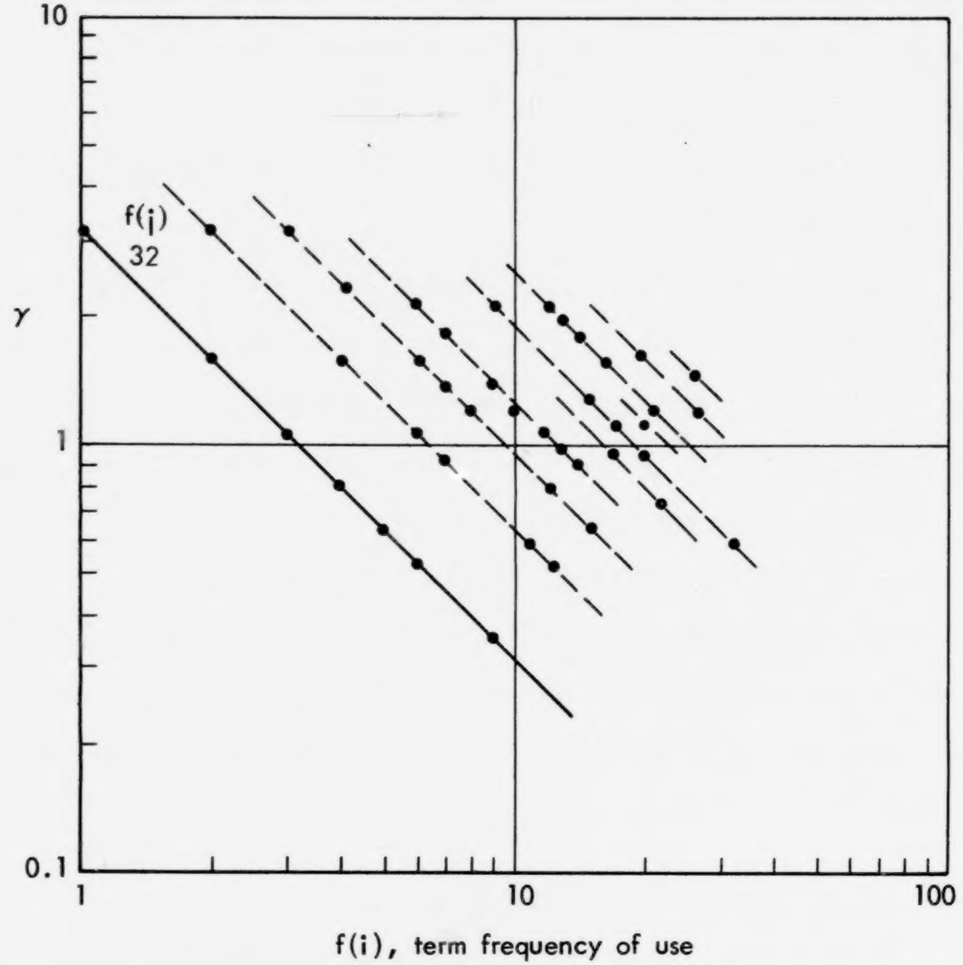


Fig.5.30 — γ factors for $f(j) = 32$ and $1 \leq f(i) \leq 32$

a curve that is an integer multiple of the lower bound value. It is always the case that the theoretical minimum value of γ is the lower bound, and that whenever there is a difference between the lower bound and the actual, the actual value is always an integer multiple of the lower bound. For $\gamma \le 5$, the dispersion of γ values is small, and increases for $5 \le f(i) \le 32$.

In an attempt to assess the distribution of the γ-factor values, plots of the cumulative distribution of occurrence versus the ratio of γ actual to γ theoretical minimum were prepared,* and are presented in Figs. 5.31 to 5.37. For terms with a high frequency of use, it is necessary to introduce a weighting factor, which as shown in the next section is a stable and well behaved factor. At this point, sufficient evidence has been accumulated (Table 5.5, and Figs. 5.15 to 5.37) to satisfy the hypothesis that the term-term co-occurrences are definable as a function of the term frequencies of use and are directly proportionate to that factor.

5.4 THE RETRIEVAL QUANTITY MEASURE

As described previously, the Retrieval Quantity (R_q) measure indicates the quantity of documents (references) that are output by a DRS in response to a formal inquiry. The purpose of this section is to develop an operational form of such a measure, and to test the measure with a set of actual inquiries on an operational system.

The procedure fo· predicting R_q for an inquiry entails several steps:

---

*Note this analysis is restricted to TXT(i,j) > 0.

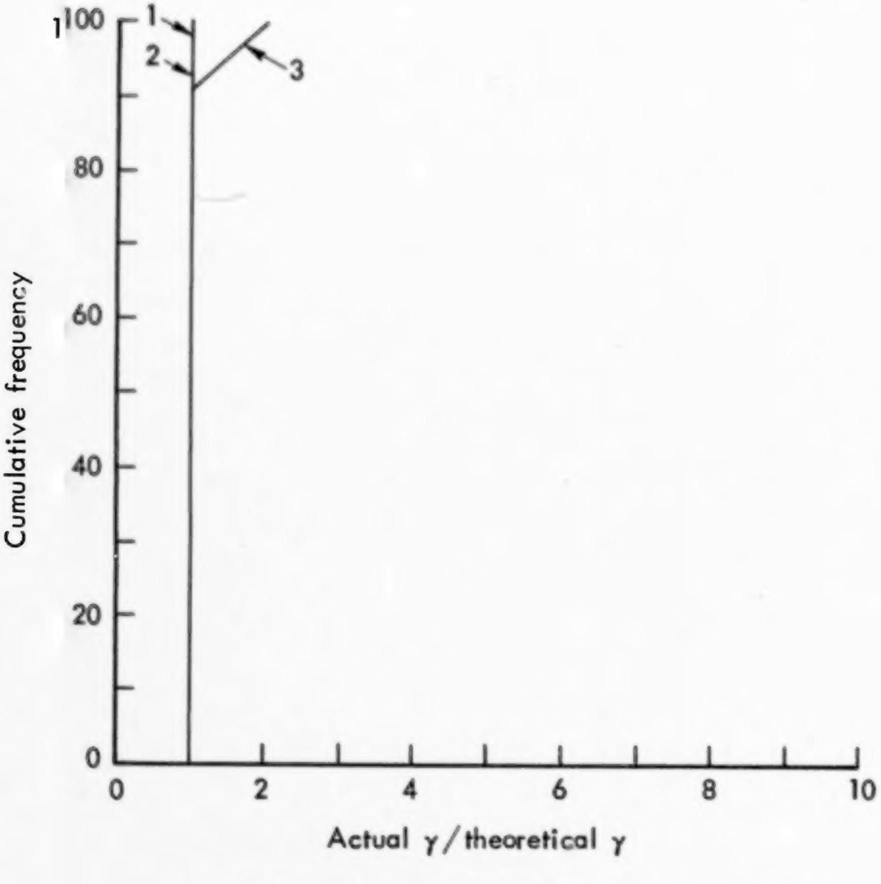


Fig. 5.31 — Cumulative frequency of the ratio of actual γ to theoretical γ for terms with frequency of use of 3 co-occurring with terms with frequency of use of 1 to 3

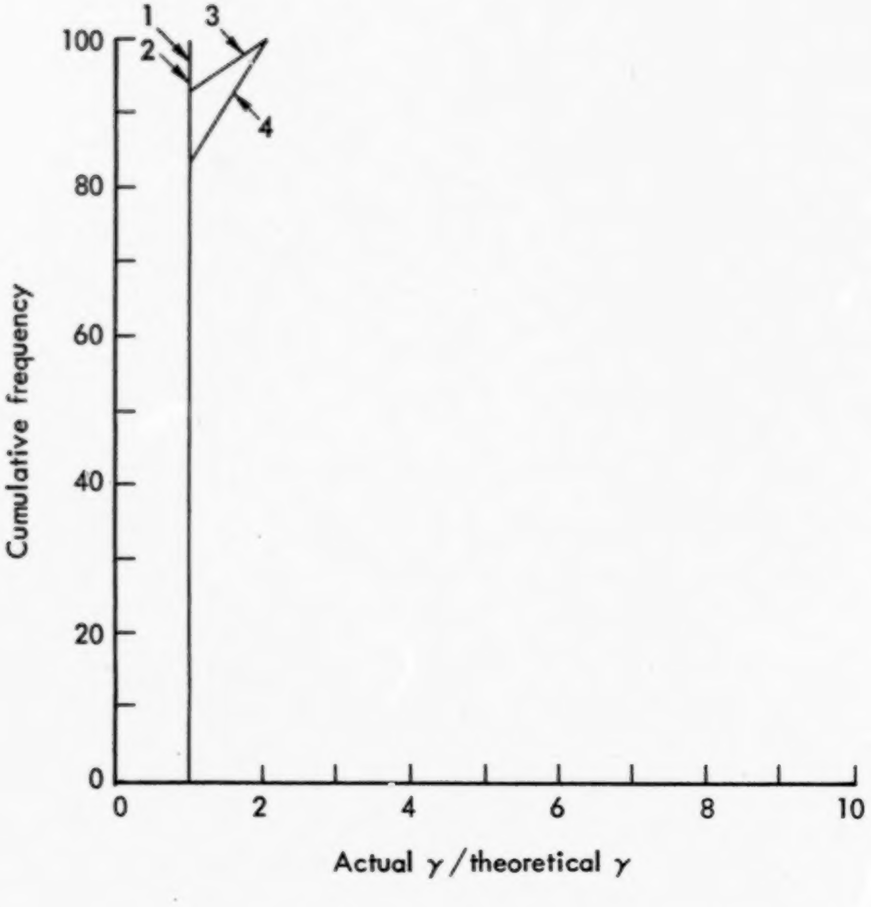


Fig. 5.32—Cumulative frequency of the ratio of actual γ to theoretical γ for terms with frequency of use of 4 co-occurring with terms with frequency of use of 1 to 4

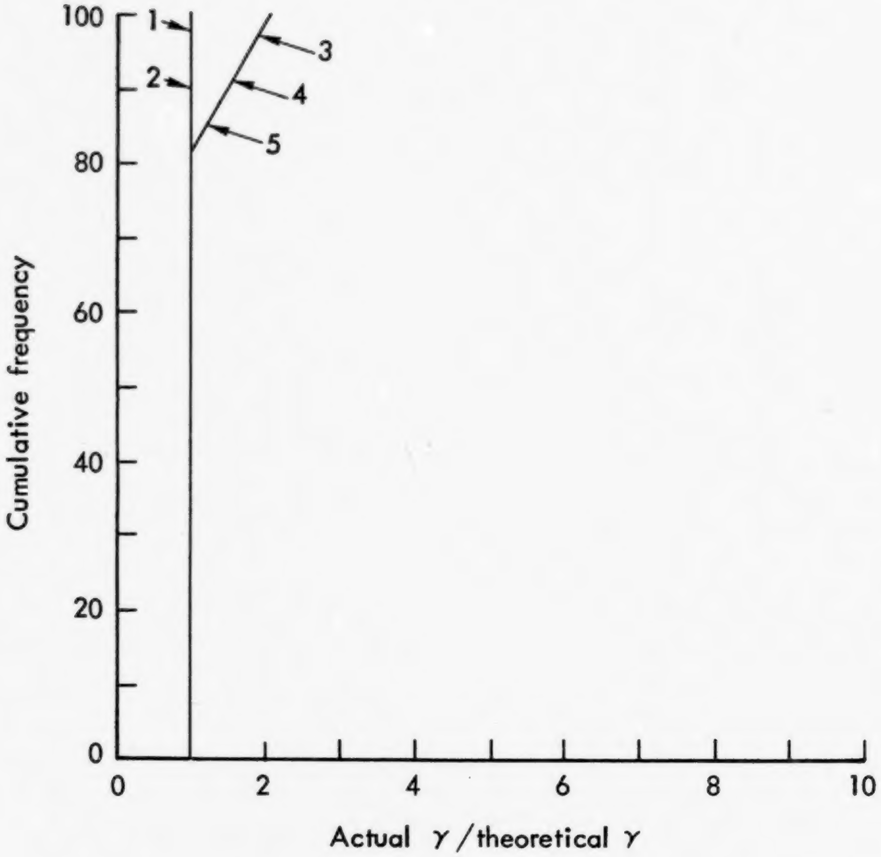


Fig. 5.33— Cumulative frequency of the ratio of actual γ to theoretical γ for terms with frequency of use of 5 co-occurring with terms with frequency of use of 1 to 5

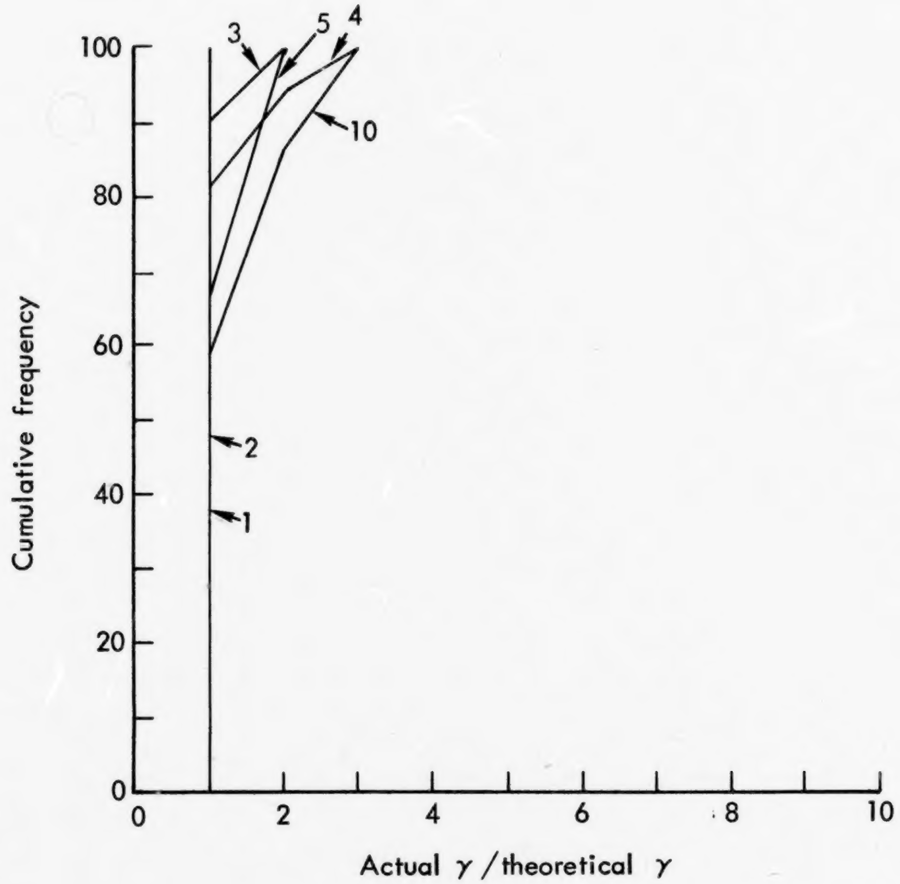


Fig. 5.34 — Cumulative frequency of the ratio of actual γ to theoretical γ for terms with frequency of use of 10 co-occurring with terms with frequency of use of 1 to 10

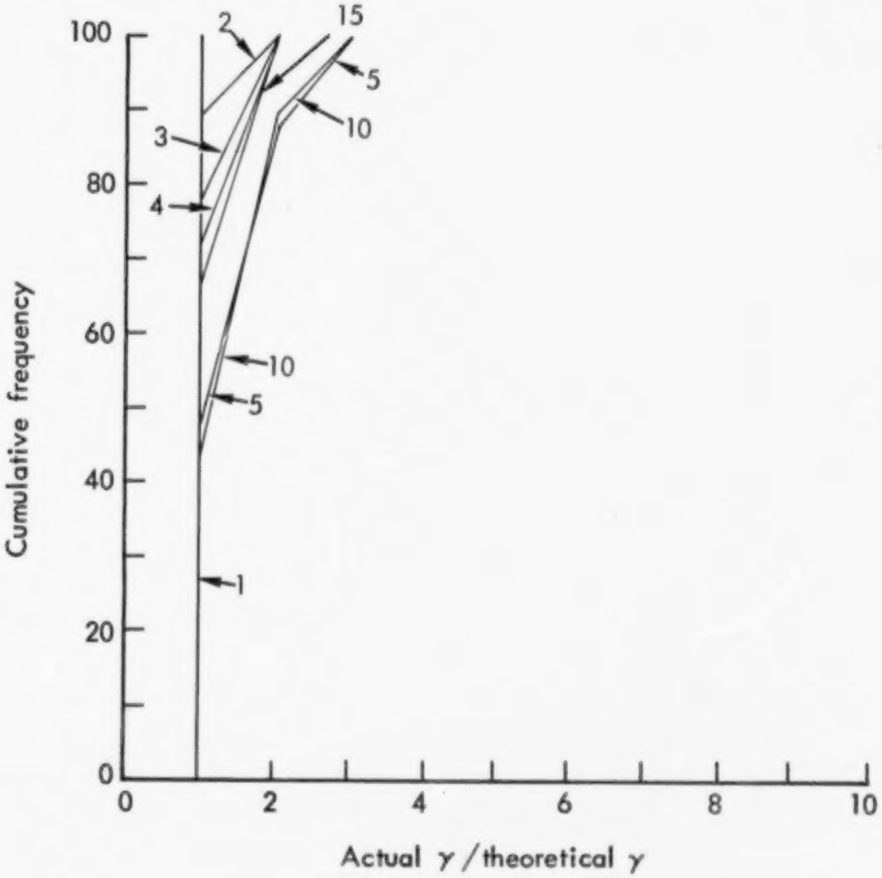


Fig.5.35 — Cumulative frequency of the ratio of actual γ to theoretical γ for terms with frequency of use of 15 co-occurring with terms with frequency of use of 1 to 15

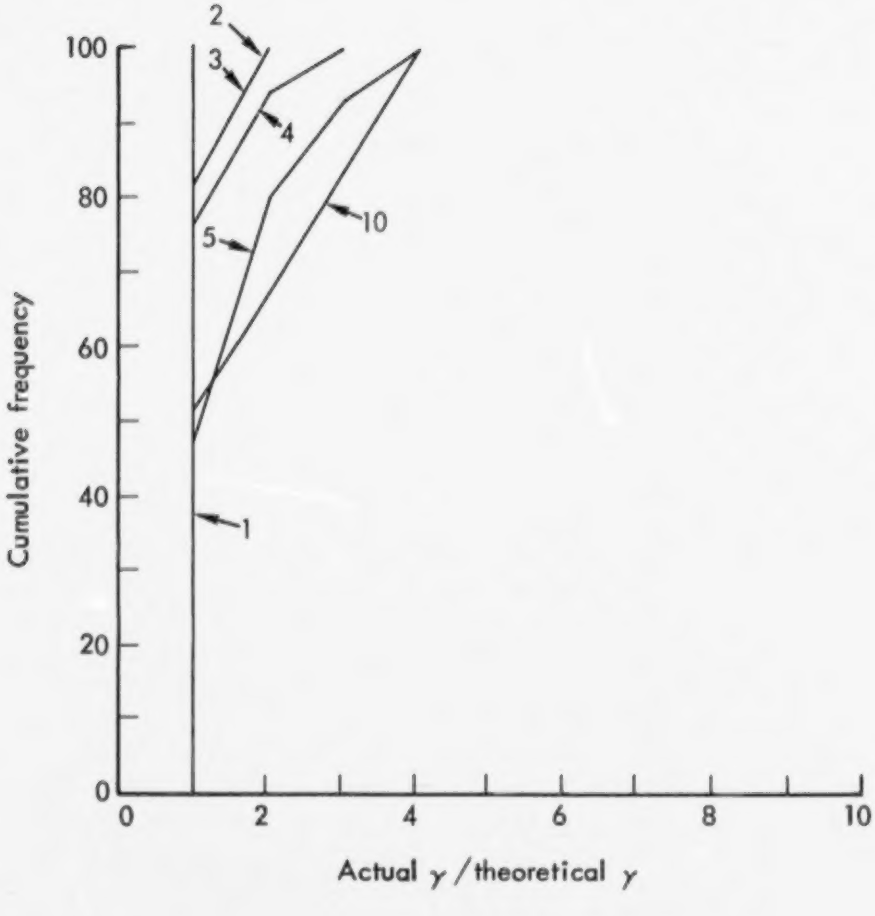


Fig. 5.36 — Cumulative frequency of the ratio of actual γ to theoretical γ for terms with frequency of use of 20 co-occurring terms with frequency of use of 1 to 10

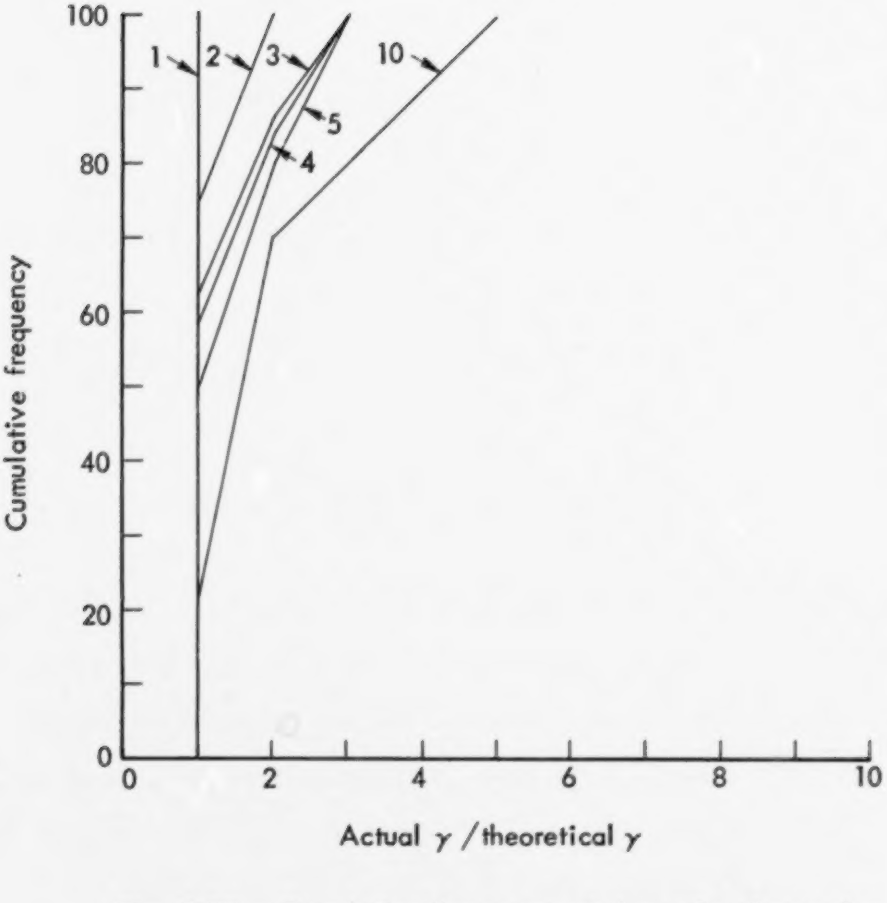


Fig. 5.37—Cumulative frequency of the ratio of actual γ to theoretical γ for terms with frequency of use of 32 co-occurring terms with frequency of use of 1 to 10

(1) construction of the formal inquiry (from the user request)

(2) application of the term co-occurrence factor -- γ

(3) determination of R_q

Step (1) has been discussed in Chapter 4, and steps (2) and (3) will be analyzed in this section.

5.4.1 Application of the Term Co-Occurrence Factor, γ, and Determination of R_q

Step (2) involves the application of γ to the explicit conjunctive arguments, and implicit intersections of the disjunctive arguments in the inquiries. Taking a simple example such as $T_1 \cdot T_2$, for which R_q is the term co-occurrence value (TXT(1,2)), the lower bound estimate of R_q is:

$$R_q = \gamma \frac{f(1) \cdot f(2)}{D}$$

γ is found by using the appropriate plot of γ and the term frequencies of use (e.g., plots like Figs. 5.31 to 5.37), and the variables f(1), f(2) and D are readily determined for any operational system.

A few examples will help to illustrate the R_q estimation procedure:

1. Request: Retrieve all those documents that discuss the concept of Coordinate Indexing

   Formal Inquiry: Concept and Coordinate Indexing

From Appendix C, the frequencies of use of each inquiry term, in the sample data are:

f (concept) = 7

f (Coordinate Index) = 10

D = 102

From Fig. 5.20, for f(i) = 10, $\gamma = 1.46$, and,

$$R_q = (1.46)(\frac{7 \cdot 10}{102}) = 1$$

which is exactly correct, for the sample data base.

2. Request: Retrieve all the documents that discuss classification and clumping

   Formal Inquiry: Classification and clump

From Appendix C, the frequencies of use for each term are:

f (classification) = 20
f (clump) = 5

From Fig. 5.28, for f(i) = 20, $\gamma = 1.02$ and the theoretical lower bound estimate of R_q is:

$$R_q = (1.02)(\frac{20 \cdot 5}{102}) = 1$$

which is less than the actual number (3) of documents described by the two terms, in the sample data base.

These examples show that for combinations of low frequency of use terms the lower bound theoretical γ-factor leads to accurate R_q estimates, but tends to diverge from TXT(i,j) as f(i) and/or f(j) increases. However, when the lower bound γ value causes the R_q estimate to be less than the actual value, the difference or correction is always an integer multiple of γ.

One way to correct for the underestimation for large f(i) is to employ a simple weighting scheme. That is, to apply weights (probabilities) to integer multiples of γ lower-bound, with the weights reflecting the proportion or frequency of occurrence of the values of

TXT(i,j) for the terms i and j of interest. For example, the R_q estimate for a n-term conjunction would be

$$R_q = (1+\alpha_1)\left[\gamma\left(\frac{f(i)\cdot f(j)}{D}\right)\right] + \alpha_2\left[2\gamma(\cdot)\right] + \ldots + \alpha_n\left[n\gamma(\cdot)\right]$$

where $n = [f(i),f(j)]_{\text{minimum}}$, and α_i can be estimated from plots of the cumulative frequency of the ratio of actual to theoretical γ's, as in Figs. 5.31 to 5.37, or from the cumulative distribution of the values of term co-occurrences, such as in Fig. 5.38, or the density distribution of the values of the term co-occurrence, as in Figs. 5.39 and 5.40.

For example 2 above, the α_i corrections are determined from Fig. 5.40, for f(i) = 20 and f(j) = 5;

$$\alpha_1 = 0.51$$
$$\alpha_2 = 0.21$$
$$\alpha_3 = 0.10$$
$$\alpha_4 = 0.06$$
$$\alpha_5 = 0.04$$

The R_q estimate is now:

$$R_q' = \left[1+(.51)(1) + (.21)(2) + (.1)(3) + (.06)(4) + (.04)(5)\right] \times \left[1.02\left(\frac{20\cdot 5}{102}\right)\right] = 2.72$$

which is a much better estimate of the actual value of 3. The distribution of α_i's is quite stable, and in Section 5.4.2 they are incorporated into the γ versus f(i) plot (see Fig. 5.42).

Unlike the above simple examples, most requests are a string of conjunctively and disjunctively related terms, and in general the string

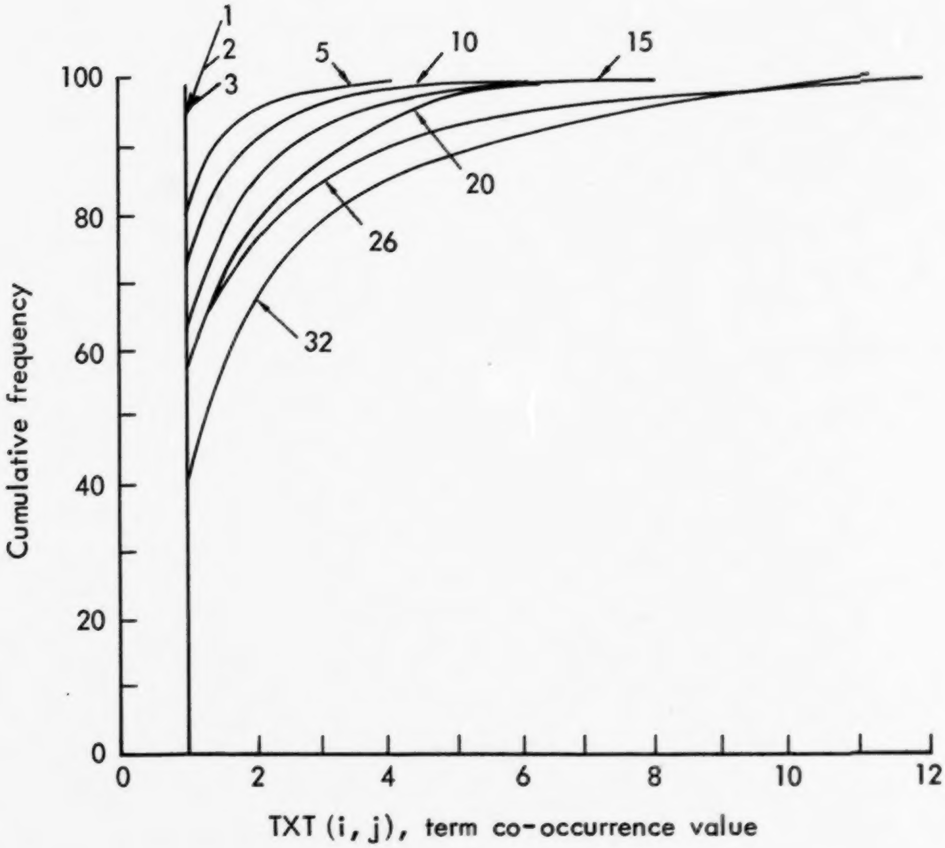


Fig.5.38 — Cumulative frequency of occurrence of TXT(i,j) for terms with f(i) = 1,2,3,5,10,15,20,26 & 32 and 1 f(j) 32; only non-zero co-occurrences plotted

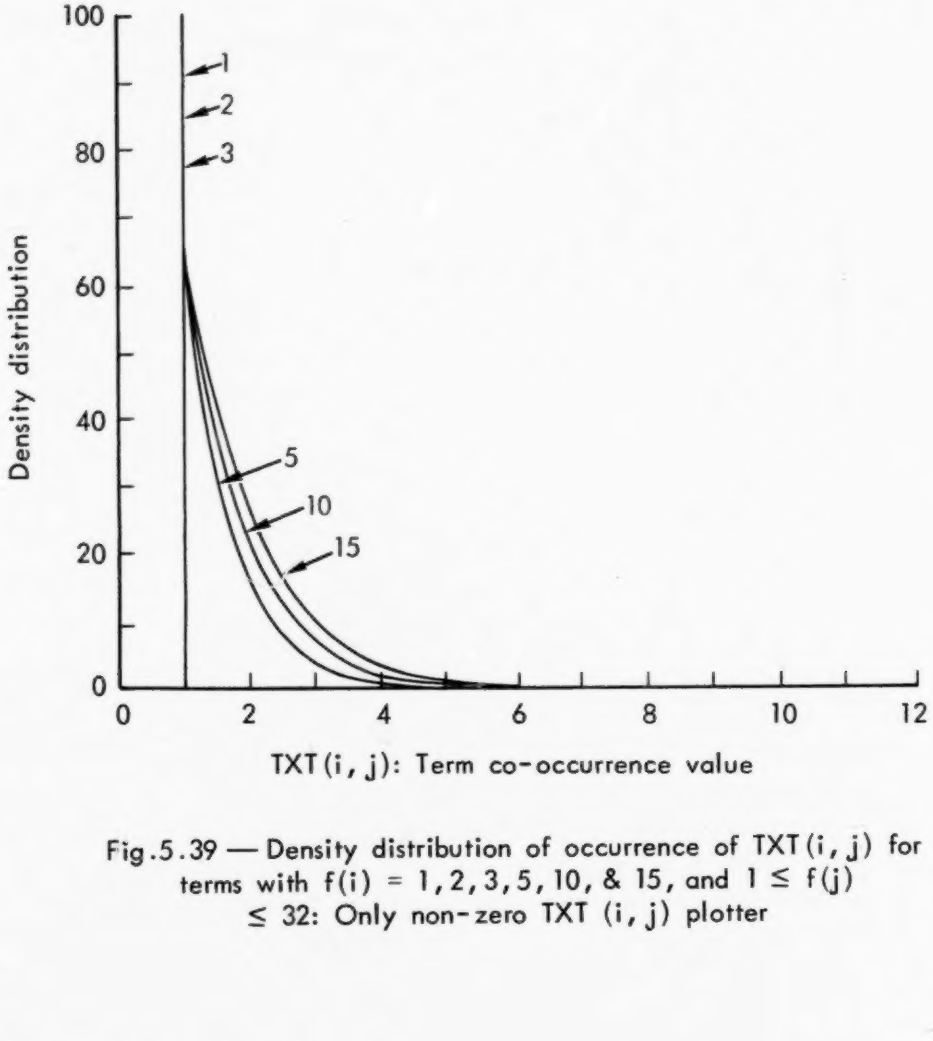


Fig.5.39 — Density distribution of occurrence of TXT(i, j) for terms with f(i) = 1, 2, 3, 5, 10, & 15, and 1 ≤ f(j) ≤ 32: Only non-zero TXT (i, j) plotter

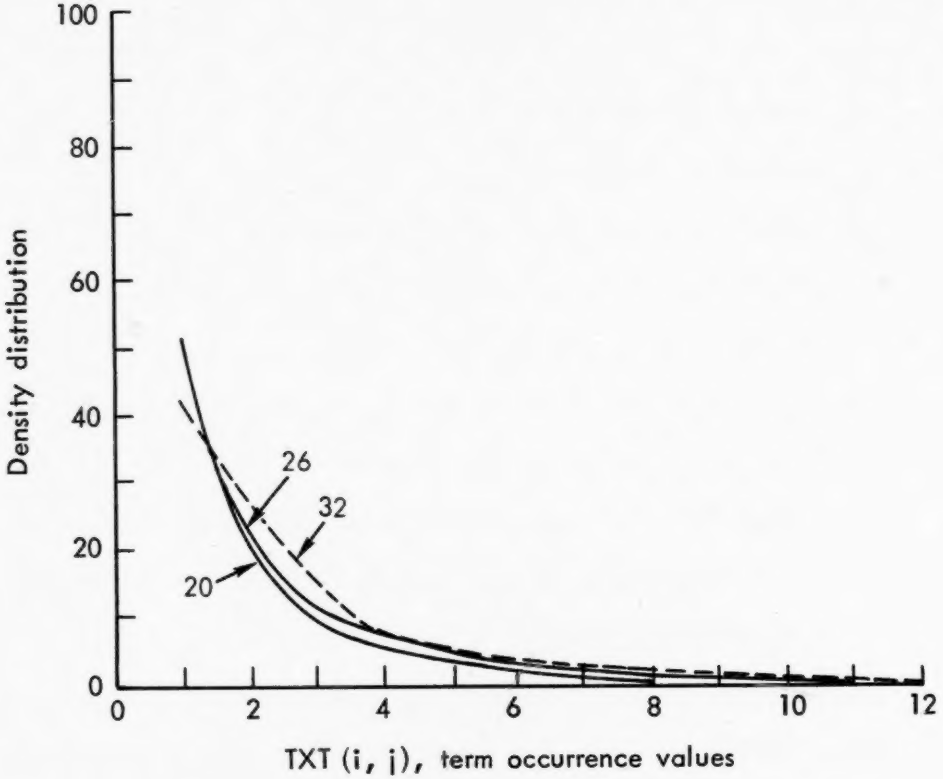


Fig.5.40 — Density distribution of occurrence of TXT (i, j) for terms with f(i) = 20, 26 & 32, and ≤ f(j) ≤ 32: only for non-zero TXT (i, j)

will contain more than two terms. When more than two terms are included in an inquiry, the estimation of R_q requires an iterative procedure. For example, consider the following inquiry:

$$T_1 \text{ and } T_2 \text{ and } T_3 \text{ and } \ldots \text{ and } T_n$$

To estimate R_q one must:

(1) determine $f(T_1),\ldots,f(T_n)$

(2) determine γ for $f(T_1)$ and $f(T_2)$, and the theoretical value of $TXT(T_1,T_2)$ by $\gamma \cdot (f(T_1) \cdot f(T_2)/D)$; this value is the intersect of T_1 and T_2

(3) call the intersect of T_1 and T_2, T_1' and determine the intersect of T_1' and T_3, as per step (2)

(4) repeat steps (2) and (3) until the intersect of T_{n-1}' and T_n is determined; this is the R_q estimate for the n-term conjunctive series

In the event that a request contains one or more disjunctions, the above iterative procedure is modified as follows. Consider an inquiry of the form:

$$(T_1 \text{ or } T_2) \text{ AND } (T_3 \text{ or } T_4)$$

To estimate R_q, recall that

$$R_q(T_1+T_2) = f(T_1) + f(T_2) - TXT(T_1,T_2)$$

and incorporate this relationship in the iterative procedure:

(1) determine $f(i)$, for $i=T_1$, T_2, T_3 and T_4

(2) determine γ for $f(T_1)$ AND $f(T_2)$, and the theoretical value of

$TXT(T_1,T_2)$ by $\gamma\cdot(f(T_1)\cdot f(T_2)/D)$; this is the intersect of T_1 and T_2, and therefore $T_1' = f(T_1) + f(T_2) - \gamma\,\frac{f(T_1)\cdot f(T_2)}{D}$

(3) repeat step (2) for all other disjunctive pairs.

(4) when all the disjunctive groups have been reduced to their "net" respective T_i''s, the remaining expression is simply a conjunctive series and the R_q estimate is determined as for the previous example.

At times an inquiry will contain an explicit negation of a term, such as in the following example:

$$T_1 \text{ AND NOT } T_2$$

To estimate R_q, an additional modification of the above procedure is required. Recall that,

$$R_q(T_1-T_2) = f(T_1) - TXT(1,2)$$

yields the net R_q. Therefore, for those clauses in which there is a negated term, the above relationship is determined, and the resulting net T_i' is used to compute the remaining conjunctions and/or disjunctions of terms.

Having established an iterative procedure to estimate quantity output for complex inquiries, the next step is the evaluation of the R_q estimation process.

5.4.2 Testing the R_q Estimate

The data and illustrations presented thus far reflect the sample data, and it is necessary to extend the findings to the test system to evaluate the R_q estimate. In order to do this, certain logical properties of the relationship

$$TXT(i,j) = \gamma \frac{f(i)\cdot f(j)}{D}$$

must be established.

As demonstrated, the above relationship is linear with slope of -1 in log-log space.* Further, for any DRS, all curves for any combination of f(i) and f(j) are derivable from the theoretical curve for f(i) = 1 and $1 \leq f(j) \leq D$. To show this, the first step is to determine the intercept for the curve f(i) = 1 and $1 \leq f(j) \leq D$.

The ordinal intercept for f(i) = 1 and $1 \leq f(i) \leq D$ is defined at f(i) = 1 and f(j) = 1, which yields

$$TXT(i,j) = 1 = \gamma \frac{(1)(1)}{D}$$

or

$$\gamma = D$$

which is the value of the intercept on the γ-axis. The intercept on the f(i) axis for the curve f(i) = 1, $1 \leq f(j) \leq D$ can be determined in a similar manner. Setting f(i) = 1, and f(j) = D yields

$$TXT(i,j) = 1 = \gamma \frac{(1)(D)}{D}$$

or

$$\gamma = 1.$$

Therefore, all one needs to know to establish the value of the intercepts for the curve f(i) = 1 and $1 \leq f(j) \leq D$, is the size D of the system corpus, and that the term usage versus rank distribution is approximated by the MEZ canonical form.

---

*For the theoretical lower bound.

The curve just determined is the lower and upper bound for all values of γ for terms with $f(i) = 1$, and $1 \leq f(j) \leq D$. In addition, this curve is the upper bound on the values of γ for all other γ for any combinations of term frequency; that is, for

$$1 \leq f(i) \leq D$$

$$1 \leq f(j) \leq D$$

Further, on the basis of the above curve for $f(i) = 1$, and $1 \leq f(j) \leq D$ the theoretical lower bound values of γ for all other combinations of $f(i)$ and $f(j)$ can be determined. The procedure to determine these lower bound curves is illustrated in Fig. 5.41, for the test corpus with $D = 416$, and $f(j) = 10$, and consists of the following steps:

(1) locate $f(j) = 10$, on the abscissa (point I in Fig. 5.41).

(2) follow the vertical line up to the intersection (point II) with the line for $f(i) = 1$, $1 \leq f(j) \leq 416$.

(3) follow the horizontal to the ordinate intercept (point III), which gives the value of γ for $f(i) = 1$ and $f(j) = 10$.

(4) trace the 45° line, with slope -1, to its intercept with the abscissa, at $\gamma=1$ (point IV)

The resulting line between points III and IV, and extrapolated beyond, is the theoretical lower bound for γ for $f(j) = 10$, and $1 \leq f(i) \leq D'$; where $D' \leq 416 - f(i)$. That is, if one were to estimate the intersect, TXT(i,j) of two terms with $f(i) = 10$ and $f(j) = 416$, respectively, it is clear that TXT(i,j) = 10, by definition. Therefore, in order that the theoretical lower bound curve satisfy that

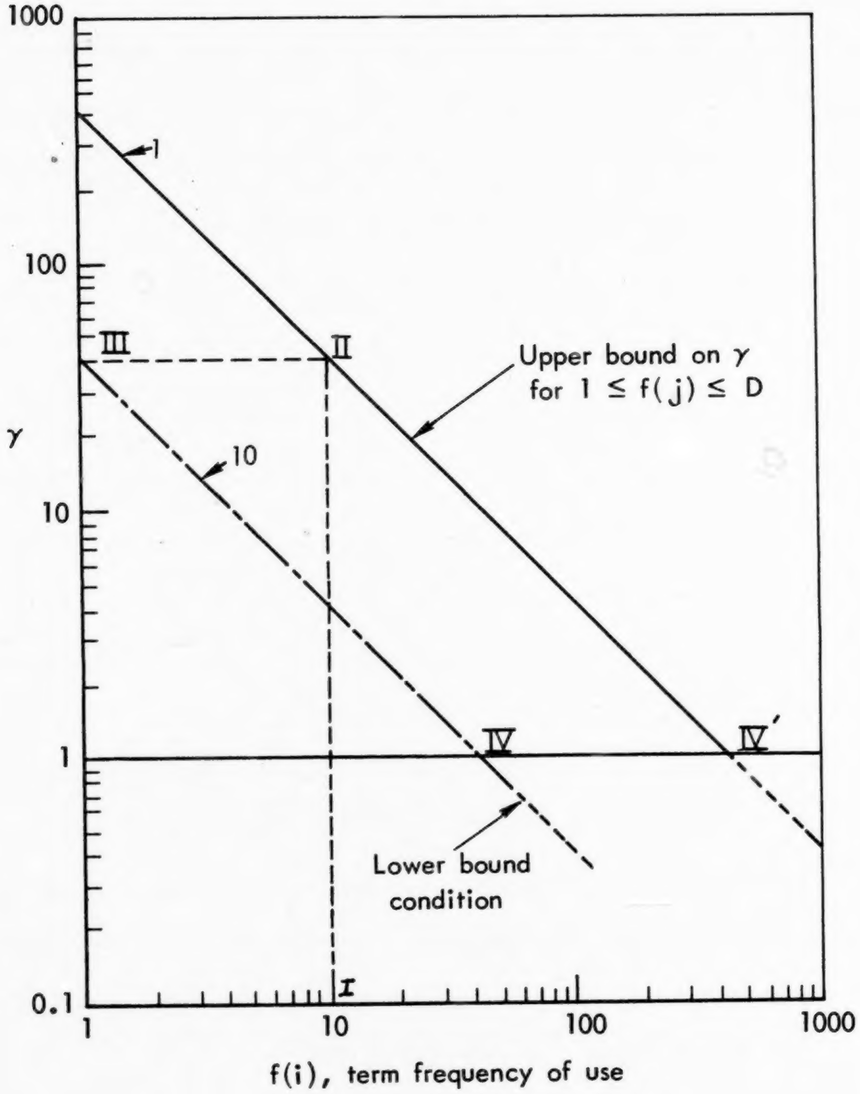


Fig.5.41 — γ factor versus f(i) for ILR document retrieval system, stage II

condition, it must be asymptotic to the line for $\gamma=1$, and intercept that line in the vicinity of $f(j) = 416$.

Given the basis for constructing the theoretical envelope, and its bounds, of γ values, the next step is to determine the best estimate values of the γ factor between the upper and lower bounds, for the test system. The best estimate values of γ can be determined using the following assumptions about -- and properties of -- coordinate index DRSs.

(1) the sample data base is representative of the parent or test system, and the divergence data indicated in Figs. 5.31 to 5.40 can be extrapolated to the value of $f(i)_{max}$ in the test system.

(2) the upper bound of the γ-curves for any term is defined by the curve of slope (-1) for $f(i) = 1$ and $1 \le f(j) \le D$, in log-log space.

(3) the lower bound of the γ-curve for any term j, is defined by the curve, with an ordinal intercept defined by the intersection of $f(j)$ with the curve for $f(i) = 1$, $1 \le f(j) \le D$, and an asymptote to $\gamma=1$ in the vicinity of D', where $D' = D - f(j)$.

(4) for any two terms, the value of the γ factor must be the same, regardless of the sequence of determination; that is, the curves must possess a symmetry such that

$$\gamma_{f(i),f(j)} = \gamma_{f(j),f(i)}.$$

This property follows from the fact that TXT is symmetric; i.e., $TXT(i,j) = TXT(j,i)$.

Using the above assumptions and properties, the best estimate γ factor curves for the test system were derived, and are presented in Fig. 5.42. From property (3), one would expect that the test system γ curves would be asymptotic to the line of slope $\gamma=1$ for $f(i) \approx D$. However, there is instead an apparent convergence of curves at $3 \leq \gamma \leq 5$, for high $f(i)$. Since the data sample had very few points in the range of $30 < f(i) < D$, it was not possible to analyze this characteristic in depth. However, it is likely that the reason for this property is that the test system is small ($D \approx 400$ and $T \approx 400$) and as the product of $f(i) \cdot f(j)$ approaches or exceeds D, the intersection of the two terms is going to be substantial, and hence the convergence of γ-curves for high $f(i)$ (but $<< D$) at $\gamma > 1$.

In order to evaluate the R_q estimation process, based on the γ-curves in Fig. 5.42, a set of 15 requests of various content was generated. The requests are considered to be typical and corpus subject related, and are not based on the descriptions of any one document or set of documents.* The test inquiries are listed in Table 5.6.

The R_q values, both estimated and actual, for each inquiry were determined for direct match searches and are reported in Table 5.7.** The estimated R_q values are, for all inquiries, very close to the actual R_q, and clearly demonstrate that the Retrieval Quantity for an operational coordinate index DRS can be accurately predicted for formal inquiries.

---

*The intent was to avoid the early Cranfield (see Ref. 130) or "Moore's" type inquiry, in which requests are generated from document descriptor sets. Such inquiries test the system retrieval search linkages, but are certainly not representative of the typical user request.

**Some sample computations are included in Appendix D.

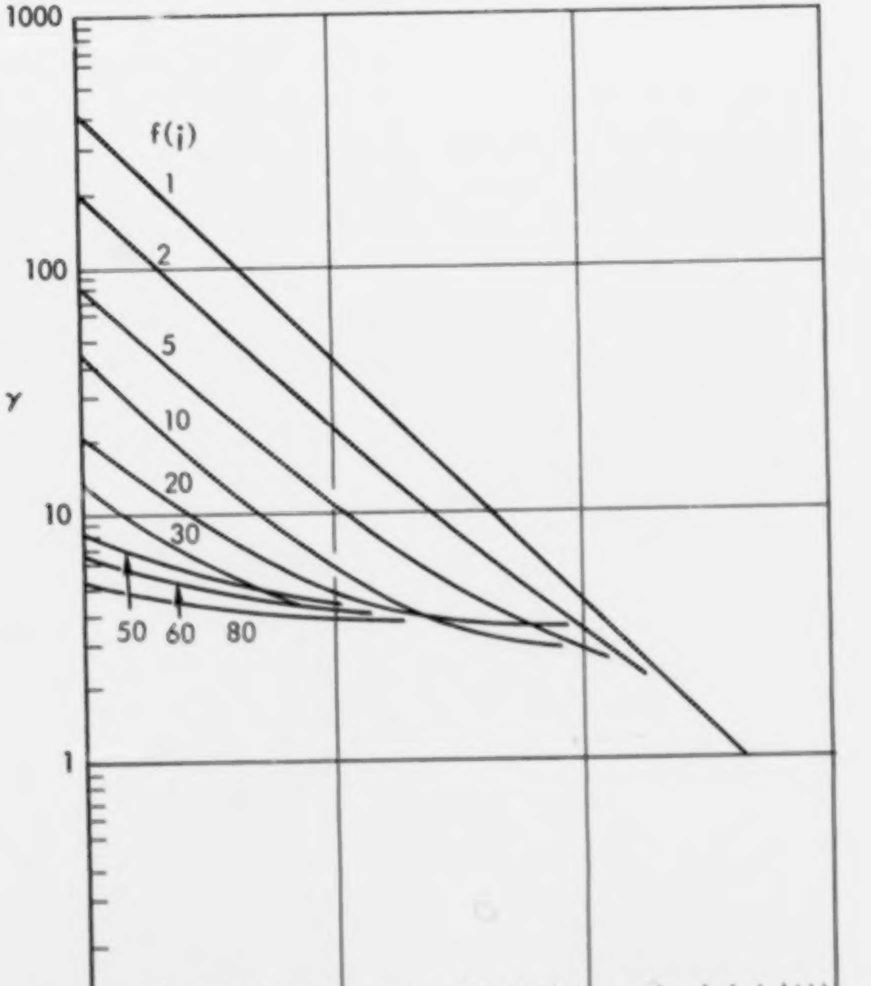


Fig.5.42 — Adjusted γ factors for ILR test system

Table 5.6

TEST INQUIRIES

Inquiry	Form	Term Frequency
1. Auto. indexing and auto. abstracting and (theory or analysis or experiment) and not manual indexing	$T_1 \cdot T_2 \cdot (T_3+T_4+T_5) \cdot T_6$	$f(T_1) = 21$ $f(T_2) = 11$ $f(T_3) = 20$ $f(T_4) = 53$ $f(T_5) = 44$ $f(T_6) = 2$
2. Comp. linguistics and syntax and semantic	$T_1 \cdot T_2 \cdot T_3$	$f(T_1) = 6$ $f(T_2) = 28$ $f(T_3) = 41$
3. Natural language and (auto. indexing or auto abstracting) and experiments	$T_1 \cdot (T_2+T_3) \cdot T_4$	$f(T_1) = 38$ $f(T_2) = 27$ $f(T_3) = 11$ $f(T_4) = 44$
4. STAT association [illegible] (clump or cluster) and experiment	$T_1 \cdot (T_2+T_3) \cdot T_4$	$f(T_1) = 10$ $f(T_2) = 17$ $f(T_3) = 13$ $f(T_4) = 44$
5. Automatic and indexing and (coordinate or subject heading)	$T_1 \cdot T_2 \cdot (T_3+T_4)$	$f(T_1) = 28$ $f(T_2) = 64$ $f(T_3) = 16$ $f(T_4) = 11$
6. Measure and relevance and evaluation and (theory or performance)	$T_1 \cdot T_2 \cdot T_3 \cdot (T_4+T_5)$	$f(T_1) = 31$ $f(T_2) = 49$ $f(T_3) = 44$ $f(T_4) = 20$ $f(T_5) = 51$
7. Simulation and (retrieval or info. retrieval or document)	$T_1 \cdot (T_2+T_3+T_4)$	$f(T_1) = 5$ $f(T_2) = 63$ $f(T_3) = 84$ $f(T_4) = 78$

Table 5.6--continued

Inquiry	Form	Term Frequency
8. Theory and (documentation or info. retrieval)	$T_1 \cdot (T_2+T_3)$	$f(T_1) = 20$ $f(T_2) = 10$ $f(T_3) = 84$
9. Design and retrieval system and (on-line or real-time)	$T_1 \cdot T_2 \cdot (T_3+T_4)$	$f(T_1) = 9$ $f(T_2) = 15$ $f(T_3) = 3$ $f(T_4) = 1$
10. Design and automatic and retrieval system	$T_1 \cdot T_2 \cdot T_3$	$f(T_1) = 9$ $f(T_2) = 28$ $f(T_3) = 15$
11. Computer and education and (design or evaluation)	$T_1 \cdot T_2 \cdot (T_3+T_4)$	$f(T_1) = 69$ $f(T_2) = 15$ $f(T_3) = 9$ $f(T_4) = 44$
12. Question and evaluation and (Boolean or logical)	$T_1 \cdot T_2 \cdot (T_3+T_4)$	$f(T_1) = 33$ $f(T_2) = 44$ $f(T_3) = 13$ $f(T_4) = 4$
13. Depth-of-indexing and (evaluation or analysis)	$T_1 \cdot (T_2 \cdot T_3)$	$f(T_1) = 8$ $f(T_2) = 44$ $f(T_3) = 53$
14. Natural language and translation	$T_1 \cdot T_2$	$f(T_1) = 38$ $f(T_2) = 31$
15. Abstracting and centers and controlled	$T_1 \cdot T_2 \cdot T_3$	$f(T_1) = 13$ $f(T_2) = 5$ $f(T_3) = 3$

Table 5.7

COMPARISON OF ACTUAL AND ESTIMATED
Rq FOR DIRECT MATCH SEARCHES

Inquiry	Rq-Actual	Rq-Estimate
1	2	1-2[a]
2	2	1-2
3	2	3-4
4	0	1-2
5	2	4
6	0	3
7	2	3
8	13	15
9	9	0-1
10	1	1-2
11	3	4
12	1	2-3
13	6	5-6
14	12	12
15	1	0-1

[a] The Rq estimate is frequently a non-integer value and the ranges indicated are integer bounds.

5.5 THE LIKELIHOOD OF NON-ZERO TERM-TERM CO-OCCURRENCES

The analysis and results presented thus far have implicitly assumed that the probability of term-term co-occurrences for terms with f(i), f(j) > 0 (for actual inquiry combinations for a homogeneous corpus) is significantly greater than zero. Thus the γ factors presented in Fig. 5.42 can be viewed as the values to estimate TXT(i,j), given that f(i), f(j) > 0 and that terms i and j do indeed co-occur. Since the DXT matrix is usually very sparse (for the test data sample approximately 95 percent of the cells are zero), and also that the TXT matrix is usually sparse* (for the test data sample, approximately 82 percent of the cells are zero), some insight into the behavior of

$$P(TXT(i,j) \mid f(i),\ f(j) > 0)$$

as a function of f(i), f(j), and the number of terms with the same frequency of use is desired..

The theoretical probability, based on independent term usage, that the co-occurrence of two terms is greater than zero, given that each term has a frequency of use greater than zero, can be determined as follows:

Given: D documents $= \{d\}$
T terms (active) $= \{t\}$

---

*It can be shown that the sparcity of TXT is always less than or equal to the sparcity of DXT; where $TXT = (DXT)^T(DXT)$ and $DXT(i,j) \geq 0$ for all i,j.

let i_t = the frequency of use of term t; $1 \leq i \leq D$

j_i = the number of terms with frequency of use i; $1 \leq j_i \leq D$

(e.g., j_3 = the number of $i_t = 3$)

where:

$$\sum_{m=1}^{k} j_m = T$$

$$\sum_{t=1}^{T} i_t = \sum_{m=1}^{k} m_{j_m} = N$$

N = the total term frequency of occurrences

For this analysis, one may specify an initial distribution for $(j_1, \ldots, j_k)$, and then for all the terms {t}, to select i_t documents at random and without replacement* and use the terms to describe the document.

For computational convenience the probability of non-occurrence, $\overline{P}(TXT(t_a,t_b) = 0$ will be determined, and then the $P(TXT(t_a,t_b) > 0 = 1 - \overline{P}(\cdot)$. A general condition on $\overline{P}$ is that:

$$\overline{P} = 0 \quad \text{for} \quad i_{t_a} + i_{t_b} \geq D$$

For the case in which $i_{t_a} + i_{t_b} < D$, the simplest situation is where only one term is used i_{t_a} times and only one term i_{t_b} times; that is, $J_{i_{t_a}} = J_{i_{t_b}} = 1$. For notational convenience, let

---

*This constraint is necessary because any one term can be assigned to any one document only once.

$x = i_{t_a}$

$y = i_{t_b}$

$$\bar{P}_{x,y}(o) = \bar{P}(TXT(t_a,t_b)) = 0$$

For this case:

I. $$\bar{P}_{x,y} = \frac{D-x}{D} \cdot \frac{D-x-1}{D-1} \cdot \ldots \cdot \frac{D-x-y}{D-y}$$

$$= \frac{(D-x)!(D-y)!}{(D-x-y)!D!}$$

However, the more general condition is when there is at least one term that is used i_{t_a} times and at least one term that is used i_{t_b} times; that is, $j_x > 1$ and $j_y > 1$ and $j_x \neq j_y$.

Let X = the number of documents described by at least one of the j_x terms with frequency of use x

Y = the number of documents described by at least one of the j_y terms with frequency of use y

Given X and Y, for those terms with the same frequency of occurrence, the probability that there are no co-occurrences $Q_{X,Y}(o)$ of these terms is exactly the probability $\bar{P}_{x,y}(o)$ defined above: that is,

$$Q_{X,Y} = \bar{P}_{x,y}(o)$$

When the specific number of co-occurrences X and Y are not known, the value of P(X) and P(Y) must be determined. Under these conditions, the probability that there are no co-occurrences is defined as

$$Q(o) = \sum_{x,y} P(X)P(Y)Q_{XY}(o)$$

$$= \sum_{X=x}^{x\cdot j_x} \sum_{Y=y}^{y\cdot j_y} P(X)P(Y)Q_{XY}(o)$$

where $P(X)$ = probability that X documents are described by those j_x terms with frequency of use x.

$$= j_x \cdot \frac{\binom{X}{x}\binom{D}{X}}{\binom{D}{x}}$$

and, similarly

$$P(Y) = j_y \cdot \frac{\binom{Y}{y}\binom{D}{Y}}{\binom{D}{y}}$$

and

$$\text{II.} \quad Q(o) = \sum_{x=X}^{x\cdot j_x} \sum_{Y=y}^{y\cdot j_y} j_x\cdot j_y \frac{\binom{X}{x}\binom{Y}{y}\binom{D}{X}\binom{D}{Y}(D-x)!(D-y)!}{\binom{D}{x}\binom{D}{y}(D-x-y)!D!}$$

A special case of the above general relationship is the probability of no co-occurrence among the j_x terms with frequency x, where for $x\cdot j_x \le D$,

(1) the number of ways the event no-occurrence can occur equals

$$\binom{D}{x\cdot j_x}$$

and

(2) the number of possible events in the space D with j_x terms,

with frequency of use x, mapped onto that space equals

$$\frac{\binom{D}{x}^{j_x}}{j_x!}$$

Thus the probability of no co-occurrence for terms j_x with the same frequency of occurrence is:

$$\text{III.} \qquad P_{x \cdot j_x} = \frac{\binom{D}{x \cdot j_x}}{\frac{\binom{D}{x}^{j_x}}{j_x!}}$$

Of particular interest are the lower bound conditions or probabilities that describe the co-occurrence of terms with $f(i)_{min}$ or $i_{t_a} = 1$; that is, terms with frequency of use of one. This probability can be viewed as the threshold case because, as shown in previous sections, the co-occurrence of terms i and j with f(i) and $f(j) > 1$ is always greater than or equal to the $f(i)_{min} = 1$ case.

A plot of the theoretical probability of at least one co-occurrence for terms with f(i) or $x = 1$ with varying values of j_x ($1 \le j_x \le D$) is presented in Fig. 5.43. In the range of $j_x \simeq 12$ it is as likely to have a co-occurrence as not, for the theoretical distribution, and for any values of $j_x > 12$ the likelihood of at least one co-occurrence is very high. The probability of co-occurrence for the actual test data is, for the few points computed, greater than or equal to the theoretical case. As such, Fig. 5.43 affords a convenient lower bound estimation on the probability of at least one

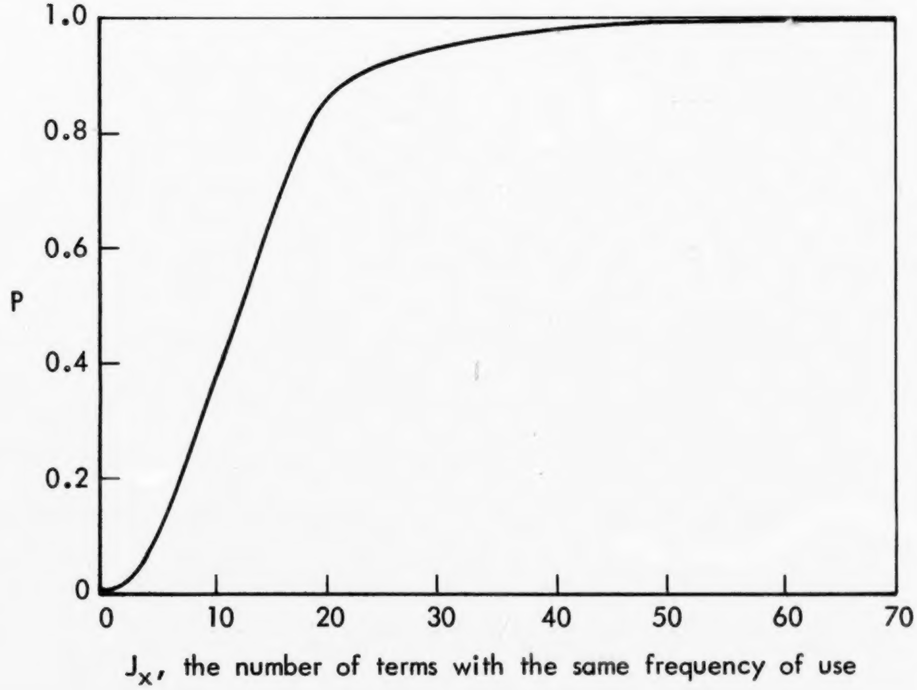


Fig. 5.43— Theoretical probability $P = 1 - (\bar{P}(TXT(i, j) = 0)$ versus $1 \leq j_x \leq 80$, for $f(i) = f(j) = 1$

co-occurrence for terms with frequency of use of 1 as a function of the number j_x of such terms.

Operationally, this means for the test sample where $j_x \approx 80$ for $x = f(i)_{min} = 1$, that one is better off assuming that a co-occurrence exists than not and at worst the R_q estimate will be off by one in a few cases.

A sample of the term-term co-occurrence for the test data is tabulated in Table 5.8. The columns are labeled in terms of the variables noted in Eq. II.

Table 5.8

TERM-TERM CO-OCCURRENCES BETWEEN TERMS WITH DIFFERENT FREQUENCY OF USE

x	j_x	y^*	j_y	$\Sigma TXT(t_a,t_b)$
1	80	1	80	83
1	↓	2	36	61
1		3	44	68
1		4	39	91
1		5	24	76
1		6	17	54
1		7	10	57
1		8	13	56
1		9	6	18
1		10	4	14
1		11	3	19
1		12	6	47
1		13	4	23
1		14	3	8
1		15	3	18
1		16	1	10
1		17	2	12
1		20	2	15
1		21	2	29
1		22	2	24
1		24	1	16
1		26	2	10
1		27	1	6
1		32	1	10

*No terms in the test data were used for f(i) = 18, 19, 23, 25, 28, 29, 30, 31.

5.6 WORD ASSOCIATION COEFFICIENTS

The relationship between the elements in the TXT matrix and the prediction functions, $\gamma(f(i) \cdot f(j)/D)$, is based on the assumption that descriptors are assigned to documents in a binary manner. That is, a term is or is not assigned as a descriptor, or in other words, the term assignment weights are 0 and 1.

In many instances, there is a need to elaborate upon an inquiry so that additional documents can be retrieved. A common technique to accomplish inquiry expansion is through word association; that is, by disjunctively incorporating new terms with those terms in the inquiry, with which they are highly correlated/associated. By necessity, these correlation relationships have non-integer values, and are derived from the TXT distribution.

In the Institute of Library Research DRS, a coefficient of association is determined for all co-occurring index terms. For purposes of processing convenience, only the four highest correlating terms are retained as association words for the base term. In the event that an inquiry is to be expanded, a disjunct is formed with the original term and its four most highly correlated terms. In general, the associated set of terms will be different for each index term, and the members of the set of associated terms can be different for any one term depending on the word association measure used.

It can be shown that the term co-occurrence factor γ can also be used to estimate word association coefficients. Following Kuhns (81), the form of a general class of coefficients of association is defined to be:

$$C_\alpha(i,j) = \frac{\alpha(i,j)}{\alpha}$$

where

$$\delta(i,j) = ITXT(i,j) - \frac{f(i)\cdot f(j)}{D}$$

A sample of the set of candidate expressions for α are listed in Table 5.9. For the derivation and rationale of these forms, and their applications, see Kuhns (81) and Maron,et al. (98), respectively.

As noted earlier,

$$TXT(i,j) = \gamma\,\frac{f(i)\cdot f(j)}{D}$$

and substituting into $C_\alpha(i,j)$, yields

$$C'_\alpha(i,j) = \frac{\frac{f(i)\cdot f(j)}{D}(\gamma-1)}{\alpha}$$

Therefore, one can estimate the coefficient of association for any two terms knowing the γ factor for the DRS.

5.7 SYSTEM GROWTH IMPACT ON RETRIEVAL QUANTITY

All operational DRSs must sustain changes in corpus collection and content, and thesaurus size in order to remain useful over time. However as the corpus and thesaurus change, particularly in size, the performance of the DRS also changes; for the same inquiry it is very possible to get different output sets from a DRS at different points in time.

In order to demonstrate the sensitivity of quantity output to changes in the system corpus and thesaurus for different search

Table 5.9

COEFFICIENTS OF ASSOCIATION PARAMETER - α (81)

Symbol	Parameter α	Description of Parameter
S	D/Z	Measure of the separation or "distance between the terms"
G	$\sqrt{f(i)\cdot f(j)}$	Measure of the angle between the vectors representing the terms
W	$\mathrm{Min}\left(f(i),\ f(j)\right)$	Measure of the conditional probability on weak evidence
R	$\mathrm{Max}\left(f(i),\ f(j)\right)$	Measure of rectangular distance between the terms
P	$\left(1 - \frac{I/T(i,\ j)}{f(i)+f(j)}\right)\cdot$ $\left(f(i)+f(j)-\frac{f(i)\cdot f(j)}{D}\right)$	Measure of the proportion overlap between the terms
L	$\sqrt{f(i)\cdot f(j)\ 1-\frac{f(i)}{D}}\ \cdot$ $\sqrt{1-\frac{f(j)}{D}}$	Measure of the linear correlation

strategies, an experiment was performed on the ILR DRS over different stages of its development. The comparative performance of the test Document Retrieval System between stage 1 and 2 is based upon a set of common questions and three word association files and direct match searches.

From the tabulated data in Table 5.10 and the plot of the measure of coefficient of word association -- G in Fig. 5.44, the dynamic property of the coefficients of association can be seen. In all cases, the S-measure produced less output as the corpus and thesaurus increased in size from stage 1 to stage 2. This is a result of both the measure and the laboratory search routine. That is, the denominator of the measure is directly proportional to any increase in corpus size, hence making the measure smaller with increasing corpus size, as the numerator increases at a much slower rate. The laboratory search routine employed also contributes to this decrease in output in that it has a default relevance threshold condition that ignores any documents that do not have a relevance value to the query, measurable in the first three significant digits. Hence any document without a relevance measure in the first three significant digits will not be retrieved.

On the other hand, the measure G provided an increase in output for all questions from stage 1 to stage 2. The W-measure provided no increase for two cases, and a slightly larger set for two cases.

It is interesting to note that the intersection of the output sets (see Table 5.10) is surprisingly small, for the same measure and same question for the two stages. Clearly, some documents that

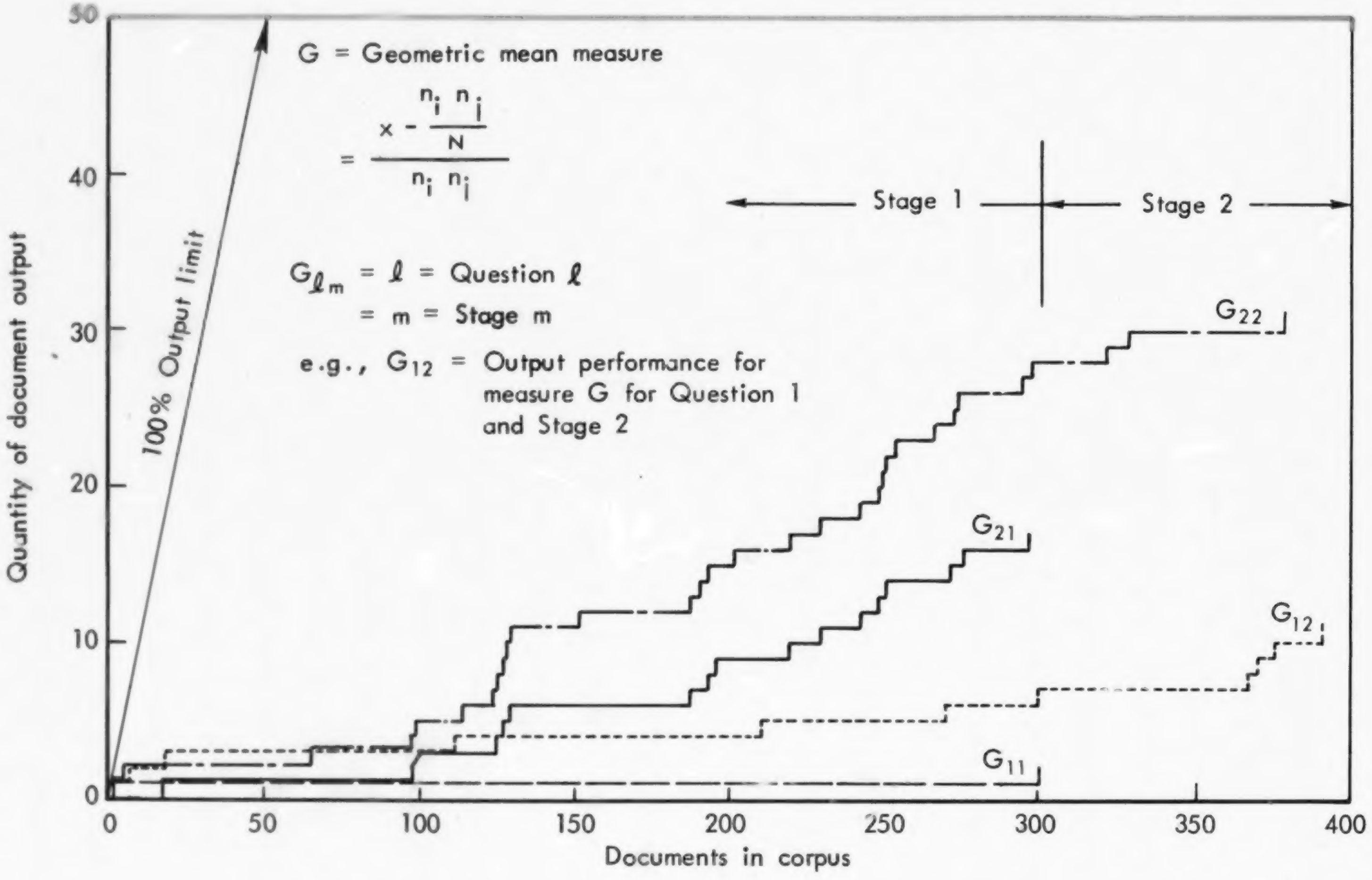


Fig.5.44 — Cumulative plot of output for coefficient of term-association G, for Stage 1 & 2 of system size, and Questions 1 & 2

Table 5.10

QUANTITY OUTPUT FOR STAGE 1 AND STAGE 2

				Cardinal Measure	
Inquiry	Coeff. of Assoc. Measure	Stage	Output Set	Intersection	Union
1	S	1	3	1	3
		2	2		
	G	1	2	2	11
		2	11		
	W	1	2	1	3
		2	2		
	Direct Match	1	1	1	1
		2	1		
2	S	1	4	1	5
		2	2		
	G	1	17	15	34
		2	32		
	W	1	8	8	14
		2	14		
	Direct Match	1	2	2	2
		2	2		
3	S	1	16	3	24
		2	11		
	G	1	1	1	23
		2	23		
	W	1	2	2	2
		2	2		
	Direct Match	1	2	2	2
		2	2		
4	S	1	4	0	4
		2	0		
	G	1	1	0	14
		2	13		
	W	1	0	0	3
		2	3		
	Direct Match	1	0	0	0
		2	0		

the system attributed as being relevant to a question in stage 1 are not being retrieved in stage 2. The cases for this difference in content of the output sets is a characteristic of the sensitivity of the different measures to system growth.

The experiment does show that the change in output performance with system growth is certainly non-linear (see Fig. 5.44). And, further, if one ignores the S-measure it can be seen from the G- and W-measures, and by examination of denominators of some of the other candidate measures in Table 5.9, that the output set will always be as large and, in the majority of cases, much larger for the same question as the system grows.

Chapter 6

CONCLUSION AND SYNTHESIS OF FINDINGS

6.1 INTRODUCTION

The purpose of the analyses in the previous chapters is to provide a basis for the development of management and design aids for DRSs, through the investigation of fundamental relationships between the components of DRSs.

The objective of this chapter is to summarize and synthesize those findings and to discuss their implications for DRS management and design.

6.2 GENERAL CONCLUSIONS

On the basis of the experiments and analysis reported in Chapter 5, it is concluded that retrieval quantity can be predicted, and that the underlying characteristics which permit the R_q estimation have potential as DRS management and design aids.

To briefly review, the findings made are believed to hold for a wide range of DRSs, such as:

Corpus size:	100 to 50,000
Thesaurus size:	300 to 13,000
Term Frequency of Use:	1 to 4,200

They are based on the detailed analysis of a representative sample DRS from this range, and consist of the following:

(1) The MEZ canonical form of $f(r) = K(r+B)^{-\alpha}$ characterizes the term-frequency-of-use versus term rank distribution for a wide range of manipulative index DRSs. The parameters K,B

and α are estimated as a function of corpus size, thesaurus size and depth of indexing.

(2) Term-term co-occurrences are not generated by random sampling from the thesaurus.

(3) The value of term-term co-occurrences is directly proportional to the function of the product of the frequencies of use of the terms, and can be predicted by the relationship

$$TXT(i,j) = \gamma\left(\frac{f(i)\cdot f(j)}{D}\right)$$

where γ is defined as a function of term frequency of use and corpus size.

(4) The Retrieval Quantity of a formal inquiry can be accurately predicted as a function of γ, term frequency of use and corpus size.

(5) For the class of coefficients of association of the form (see Kuhns (81))

$$C_{\alpha}(i,j) = \frac{\alpha(i,j)}{\alpha}$$

the numerator,

$$\delta(i,j) = TXT(i,j) - \frac{f(i)\cdot f(j)}{D}$$

can be estimated by

$$\delta'(i,j) = (\gamma-1)\left(\frac{f(i)\cdot f(j)}{D}\right)$$

(6) The probability that two terms, with frequencies of use greater than or equal to one, will co-occur is definable by an ordered family of curves with an upper and lower bound as indicated in Fig. 6.1. Each curve is a function of the frequencies of use of the two terms, the number of terms with the same frequency of use, and the size of the corpus.

(7) Terms with the same frequency of occurrence, have similar DRS statistical properties; that is, the distribution of the number and value of their term-term co-occurrences are approximately the same.

(8) The impact of DRS corpus and thesaurus growth on retrieval quantity can be predicted.

6.3 MANAGEMENT AND DESIGN AIDS

The management of a DRS entails cost/benefit analysis of system operations and plans, measuring system performance for different tasks, and controlling the system processes. It is not the intent to delve into a discourse on DRS performance evaluation, but rather to describe how the findings (summarized above) can be used to aid in some aspects of DRS management and design.

(1) Tuning Inquiries. By estimating R_q for an initial inquiry, the grammatical combinations and/or number of terms can be modified to yield different expected R_q's. Through this pre-processing exercise the DRS user can adjust inquiries to retrieve a more preferred quantity of references. In this way the marginal effect on quantity output of adding or deleting a term of a certain frequency of use, and creating

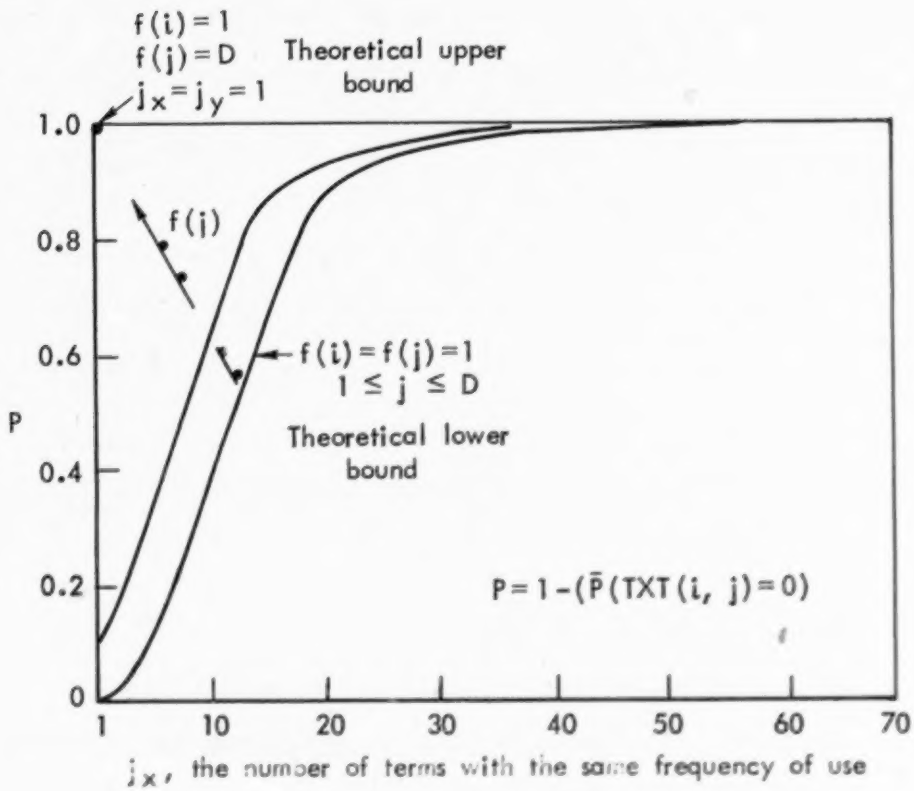


Fig. 6.1 — Theoretical family of curves defining the lower bound of the probability of co-occurrence of two terms with $f(j) = 1$, $1 \leq f(j) \leq D$, and $1 \leq j_x \leq D$, $j_y = 1$

different logical combinations can be estimated.

By employing such a "tuning" measure it is quite likely that the DRS users will find the system more understandable and convenient, and management can reduce the potential number of user disappointments in system responses.

(2) Predicting and Monitoring the Impact of System Growth. As the system corpus and thesaurus change over time, both the quality and quantity of the system output will also change, for a constant set of inquiries. The R_q measure can be used to estimate the impact of corpus and thesaurus change on the system output quantity. The most straightforward application is to determine the set of γ factors for an operational DRS with a specified corpus and thesaurus size, and then as D is increased to project a proportionate increase in the γ-factor bounds. The new γs can be used to estimate the changes in R_q, for a specific inquiry. Using the R_q measure in this way provides some insight into the dynamic characteristics of DRSs.

One could also use the R_q measure to estimate the impact on output quantity due to changes in the thesaurus with the corpus held constant. In this process, the frequency of use of the thesaurus terms would be changed, and/or new terms added. The bounds of the γ-factor would remain the same, but the likely value of γ for high f(i) would change, and the technique for estimating the new γ's is directly analogous to that used in Section 5.6, to illustrate the

$P(TXT(i,j)) > 0$ distribution.

(3) Indexing Process Modification. There are various controls that can be imposed on the indexing process, and the R_q measure can be used to estimate the effect of changes in control limits on the quantity output. For example, a manager or designer may want to:

a. Truncate the index term frequency of use distribution by specifying $f(i)_{min}$ and/or $f(i)_{max}$ limits. The impact, on quantity outputs, of changing the values of $f(i)_{min}/_{max}$ can be estimated by computing R_q at the different values, for a set of typical inquiries.

b. Limit the minimum or maximum number of terms that can be used to describe any one document. An interesting condition to investigate is to alter the "current" depth of indexing, D_E, lower and upper bounds so as to gradually approach a uniform distribution in which $D_{E_{min}} = D_{E_{max}}$. The sensitivity of the quantity output to the rate of change of the depth of indexing distribution can be estimated by the R_q measure, because the frequency of term use, $f(i)$, distribution is indirectly altered and R_q is a function of the values of $f(i)$.

c. Specify a limit or a certain distribution on the number of terms that can have the same frequency of use, j_x, over the term-rank space $\{1,\dots,D\}$. By altering the j_x value or distribution, the P_r(TXT (i,j)=V), $0 \leq V \leq (f(i),f(j))_{min}$ and $P_r(\gamma=Z)$, $\gamma_{L.B.} \leq Z \leq \gamma_{U.B.}$ probability distributions are changed, and consequently the quantity output for any one inquiry will also be modified. The impact can be estimated by R_q, because it is a function of the various γs related to the terms in the inquiries.

(4) Inquiry Processing Effort. Given a specifiable file structure and an elapsed time distribution for term lookups, the number of iterations involved in the determination of R_q can be used to estimate the average amount of time to process an inquiry. This information could be used by a DRS manager or designer to estimate certain resource requirements necessary to satisfy existing or projected user demands.

The above exemplary management applications of the R_q measure can also be viewed in the context of a design process. Combining these applications with certain canonical expressions, noted in Chapters 4 and 5, that characterize the fundamental relationships in DRSs, one can construct a hypothetic sequence of steps which illustrates their use in the design process. Further this procedure can be considered as a basis for a simulation model that would enable a designer to experiment with different parameter values and variable limits, prior to the construction of the DRS. The steps envisaged are as follows:

(1) Selection of Corpus Topic

  a. Analysis of user needs

  b. Selection of the published subject area of interest; for example, the field of Operations Research.

(2) Identification of Periodical Population and Determination of Periodical Productivity Distribution

  a. Determination of the tradeoff between number of periodicals to be collected versus the percent of the relevant literature covered, by applying Bradford's Law of Scatter (88). Kendall (75) has in fact investigated the periodical productivity distribution for Operations Research and found that if one collected the five most productive journals, 33 percent of the new articles (documents) would be captured, or the eighteen most productive journals, 50 percent of the new articles would be captured, or the 67 most productive journals would yield 75 percent of the new articles, etc.

  b. Estimation of the expected growth rate of the literature in the field, and conversely, the death or deletion rate. In most cases a sample exponential form as in Fig. 1.5 can be utilized.

(3) Estimation of the Corpus Size D

  a. From the determination of the required number of periodicals to be collected, an estimate of the initial corpus size, D, can be made.

(4) Selection of Candidate Term Frequency of Use Distributions

a. The most convenient relationship to employ is the MEZ canonical form, with the parameters K, B and α determined as in Section 5.3 that is compatible with a corpus of size D and selected average depth of indexing (e.g., D_E = 15 terms per document).

(5) Determination of the Probability of Term Co-occurrence

a. As a function of the term frequencies of use (f(i)), the size of the corpus (D), and the distribution of the number of terms with the same frequency of use (estimated as in Section 5.3.2),* the probability of two terms with frequencies of use f(i),f(j) co-occurring can be determined, as discussed in Section 5.5.

(6) Derivation of the γ-Factors for R_q

a. Based on the information determined in steps 4 and 5, the γ-factor distribution can be derived as shown in Section 5.4.1.

(7) Generate Sample Inquiries

a. A set of "typical" inquiries, from the point of view of form (and not content), can be constructed using combinations of Boolean connectors and terms with various frequencies of use as specified by the MEZ distribution.

---

*An alternative approach is to employ the Waring distribution; see Herdan (64, 65) and Jones (73) for a discussion of this distribution.

(8) Estimation of Quantity Output, R_q

a. Using the γ-factor distribution and the procedure developed in Section 5.4.1, the quantity output for the candidate inquiries can be predicted (for a direct match search strategy).

(9) Measurement of the Sensitivity of R_q to:

a. Changes in the corpus and thesaurus size

b. Changes in the MEZ parameters

c. Changes in the distribution of the number of terms with the same frequency of use

d. Changes in search strategy

The standard process of designing DRSs is considerably more art than science, with many system variables and relationships at best indirectly controlled or left to assume "natural" values by implicit default options. This process can be improved by simply taking advantage of the statistical regularities that characterize the relationship among DRS parameters. The hypothetic design sequence described above is one way in which the design process can be made more formal and accurate. Also it provides a basis for a structure within which a designer can exploit the various canonical forms that characterize the statistical stability of various DRS properties.

Chapter 7

RECOMMENDATIONS FOR ADDITIONAL RESEARCH

7.1 INTRODUCTION

There are a number of directions for future research in the area of analytic/simulation modeling of Document Storage and Retrieval Systems. Several suggestions are briefly noted in this chapter in the hope that they will provide a point of departure for one or more subsequent research efforts.

7.2 CORPUS HOMOGENEITY AND HETEROGENEITY

The DRSs investigated in this study are basically homogeneous in subject content; that is to say, the corpus is dedicated to a single subject. The ILR DRS has a homogeneous corpus and the subject is Information Science. A measure to distinguish between a homogeneous and heterogeneous corpus has yet to be developed. Also, a means of measuring the impact of more or less heterogeneity on DRS performance is needed.

Presumably, a measure could be based in part on the characteristics of the DXD matrix, which is defined by the operation

$$(\mathrm{DXT})(\mathrm{DXT})^T.$$

The DXD matrix gives the document-document association profiles, and presumably in a homogeneous corpus the majority of documents would be highly associated. The converse would hold for a heterogeneous corpus.

7.3 DISTRIBUTION OF TERMS WITH COMMON FREQUENCIES OF USE

Little, if any, control is ever exercised over the number of terms allowed to have the same frequency of occurrence, Jx. From the MEZ relationship, the Waring distribution (see Herdon (64, 65)) and Zipf's two "Laws" (see Booth (10)), there is an implied increase in Jx as the rank of the term decreases. This simply means that there will be more terms that are used infrequently than there are terms that are used frequently. The issue of interest is, what should Jx be for a specified term rank and for certain system characteristics -- D and T, and what is the impact of Jx on DRS performance.

It is clear that Jx has a marked impact on the probability of co-occurrence of terms with frequencies of use f(i), f(j). This is illustrated in Fig. 5.43, in which the theoretical lower bound of the actual P(TXT(i,j)/f(i),f(j) > 0) is plotted for f(i) = f(j) = 1 and 1 ≤ Jx ≤ D. The various formulae presented in Sec. 5.5 provide a point of departure, for any additional computations of P(TXT(i,j) = S) for a specific f(i), f(j) and Jx.

7.4 THE MEZ CANONICAL FORM

Mandelbrot (94, 95, 96), Herdan (64, 65), Zipf (153), and Leavitt (80) have investigated various term usage relationships, primarily in a text-free setting. For that unconstrained setting, the MEZ exponent α is considered always to be in the range of $1 \leq \alpha \leq 1.6$. However, the system vocabularies of DRSs are very constrained (in the predicate calculus sense), and for the test system a very good fit between the MEZ and the term frequency of use versus rank curve was possible with $\alpha = 0.9$. Clearly if one were to reduce α to zero, the frequency of use versus rank distribution would yield a uniform distribution. Intuitively then as one reduces α one constrains the "richness" of the vocabulary. Noteably, Mandelbrot (94) has observed that in children's talk (an example of constrained vocabularies of a different type) it is possible for $\alpha \leq 1$. The issues of interest are: What should α be in order that the DRS perform well, and how can one best adjust the DRS to move toward a more preferred term frequency of use situation? And, as the DRSs grow over time, what changes can be expected in the parameters K, B and α.

7.5 DEPTH OF INDEXING DISTRIBUTION

The depth of indexing distribution portrays the frequency distribution of the assignment of terms to documents. Of the systems[*] on which empirical data was available, the basic form of the distributions is very similar; in fact, sufficiently similar for one to suspect that a canonical form should exist. On the basis of a crude fit, the Beta distribution:

$$f(w,x,\phi) = \frac{(x+\phi+1)!}{x!\phi!} w^{x}(1-w)^{\phi}$$

where w is the normalized depth of indexing level defined over the finite interval $0 \leq w \leq 1$, and x and ϕ are constants. Wiederkehr (143) has developed certain forms for a modified Beta distribution in his discussion of search characteristic curves. Also, Bourne (13),

---

[*] The ILR test system and the systems investigated by Litofsky (90).

Svenonius (127), Swanson (128) and Zunde (151) have explored various aspects of the depth of indexing distribution. However, no general formulation of the expected or likely depth of indexing distribution has been developed, and just as importantly there is no established means of linking the depth of indexing characteristics with the term frequency of use distribution, and the DRS performance.

7.6 HIGHER ORDER TERM ASSOCIATIONS

The vast majority of discussions (this paper included) dealing with term-term associations just employ the first order TXT matrix relationships. As noted in Chapters 4 and 5, the elements TXT(i,j) provide the degree of association between terms i and j, which is also the first order of association. To obtain the higher order associations between two terms, one merely takes the appropriate power of the TXT matrix. That is, $(TXT)^n$ yields the n^{th} order association between the terms in the thesaurus. Salton (117) has suggested a scheme to utilize the higher order associations for expanding an initial inquiry. The procedure entails a weighting factor α, where $0 < \alpha < 1$ which causes α^n to be a monotonically decreasing function as n increases. This condition implicitly states that the lower order associations are more important than the higher order associations. Employing Salton's notion of a normalized query vector, $\bar{Q}$, one then gets the following relationship between an expanded query $\bar{Q}_E$, and the original query $\bar{Q}$:

$$\bar{Q}_E = \bar{Q}[1 + (\alpha(TXT))^1 + (\alpha(TXT))^2 + \ldots + (\alpha(TXT))^n].$$

Given that this type of relationship is valid, what are the reasonable values of α and n, and what are their effects on the performance of the DRS?

7.7 R_q MODEL EXTENSIONS

Given the basic construct of the R_q model, it is of interest to consider how the model can be extended to deal in some way with the issue of relevance.

The most logical step is to employ some means of ranking the documents by degree of inquiry term/document descriptor overlap or associative thresholds, or by the weak ordering action suggested by Cooper (35). The important procedure is to link the R_q output set with a relevance measure, which in this case would be system defined (as opposed to user judgment). Obviously, the simplest case is for a direct match search strategy in which the documents retrieved that satisfy any explicit or implied conjunction combination of terms in the inquiry would be judged the most likely relevant subset, and the documents generated by the disjunctive arguments in the inquiry less likely to be relevant. The analogous argument would hold for a word association search strategy. This elementary ranking of the output set would yield at best a binary relevance mapping on R_q, which is less discriminating than desired.

A more sophisticated approach would be to employ a probabilistic mechanism in the DXT matrix that would reflect both the fundamental indefiniteness[*] in the indexing term selection process, and the strength of the term-document assignment. Thus given a term-document relevance "weighting" one could introduce relevance thresholds in the R_q iterative procedure and potentially rank the output set. The probabilistic structures put forth by Maron and Kuhns (97) and Bryant (23) appear to be most appropriate.

7.7.1 Psychological Analogies

A rather innovative extension of the R_q model structure is to attempt to characterize the conceptual "dual" or analogous psychological process experienced by humans in searching for or processing information, by a similar model construction.[**] That is to say, there are certain regularities that characterize Document Retrieval Systems, and it is of interest to know whether these are analogous regularities that characterize the human thought process of information storage and retrieval, and, in particular, indexing and abstracting processes.

There appears to be a sound, though largely unexploited, logical basis upon which to investigate the above notion. For example, the MEZ relationship is known to characterize the work frequency of occurrence and rank distribution of a variety of languages. In fact,

---

[*] This indefiniteness arises more from a type of intrinsic uncertainty or ambiguity than from statistical variation -- a sort of "fuzzy" membership of a term to a document descriptor set (see Zadeh (151)) for a fuller discussion.

[**] Suggested by Professor F. N. Nicosia, Graduate School of Business Administration, University of California, Berkeley

Mandelbrot (94, 95, 96) (see also Brillouin (18)) derived that relationship employing the notion of the "cost" of a word as the indicator of its likelihood of use. The hypothesis is that the less costly words are used more often than the more costly, where cost is a surrogate for "effort" to use. Also, Zipf (153) presented the "law" of term frequency of use versus rank within the context of his theory on Human Behavior and the Principle of Least Effort (153). An attempt was made by Rosenberg (115) to utilize the Zipf relationship for predicting index term selection for use, but the performance of that model clearly needs to be improved before an operational construct can be developed. It would seem that a weighted Bayesian or conditioned probability structure is needed to accommodate the many degrees of semantic uncertainty and noise embedded in document discussions, human communication and indexing.

BIBLIOGRAPHY

1. A. D. Little, Inc., Centralization and Documentation, July 1963.

2. ______________, Appendices to Centralization and Documentation, July 1963.

3. Artandi, S., An Introduction to Computers in Information Science, The Scarecrow Press, Inc., Metuchen, New Jersey, 1968.

4. Baker, N. R., and R. E. Nance, "Organizational Analysis and Simulation Studies of University Libraries: A Methodological Overview," Information Storage and Retrieval, Vol. 5, 1970.

5. Barhydt, G. C., "A Comparison of Relevance Assessment by Three Types of Evaluations," American Documentation Institute: Parameters of Information Science, Spartan Books, Washington, D. C., 1964.

6. Becker, J., and R. Hayes, Information Storage and Retrieval: Tools, Elements, Theories, John Wiley, New York, 1963.

7. Bernier, C. L., "Correlative Indexes versus the Blank Sort," American Documentation, Vol. 9, 1958.

8. Beyer, W. H. (ed.), Handbook of Tables for Probability and Statistics, The Chemical Rubber Company, 1966.

9. Birkoff, G., Lattice Theory, American Mathematical Society, Providence, Rhode Island, 1966.

10. Booth, A. D., "A 'Law' of Occurrences for Words of Low Frequency," Information and Control, Vol. 10, No. 4, April 1967.

11. Borko, H., Evaluating the Effectiveness of Information Retrieval Systems, System Development Corporation, SP 909, August 1962.

12. Bourne, C., "The World's Technical Journal Literature: An Estimate of Volume, Origins, Language, Field Indexing, and Abstracting," American Documentation, April 1962.

13. __________, Methods of Information Handling, John Wiley, New York, 1966.

14. __________, "Evaluation of Indexing Systems," Annual Review of Information Science, Vol. 1, Interscience, New York, 1967.

15. __________, et. al., Requirements, Criteria, and Measures of Performance of Information Storage and Retrieval Systems, Stanford Research Institute, December 1961.

16. Bradford, S. C., Documentation, Crosby Lockwood, London, 1948.

17. Brandhorst, W. T., "Simulation of Boolean Logic Constraints Through the Use of Term Weights," American Documentation, Vol. 17, No. 3, 1966.

18. Brillouin, L., Science and Information Theory, Academic Press, New York, 1962.

19. Brookes, B. C., "The Growth, Utility, and Obsolescence of Scientific Periodical Literature," Journal of Documentation, Vol. 26, No. 4, December 1970.

20. ____________ , "The Derivation and Application of the Bradford-Zipf Distribution," Journal of Documentation, Vol. 24, No. 4, December 1968.

21. ____________ , "Obsolescence of Special Library Periodicals Sampling Errors and Utility Contours," Journal of ASIS, September/October 1970.

22. ____________ , "The Design of Cost Effective Hierarchical Information Systems," Information Storage and Retrieval, Vol. 6, 1970.

23. Bryant, E. C. (ed.), Evaluation of Document Retrieval Systems: Literature, Perspective, Measurement, Technical Papers, Westat Research, Inc., December 31, 1968.

24. ____________ , et. al., Associative Adjustments to Reduce Errors in Document Screening , Westat Research, Inc., 1967.

25. ____________ , "Modeling in Document Handling," Electronic Handling of Information, Kent and Taulbee (eds.), Academic Press, London, 1967.

26. Bush, V., "Memex Revisited," Science is Not Enough, New York, 1967.

27. _______ , "As We May Think," Atlantic Monthly, Vol. 176, July 1965.

28. Carter, L. F., et. al., National Document Handling Systems for Science and Technology, John Wiley, New York, 1967.

29. Churchman, C. W., The Systems Approach, Delacorte Press, New York, 1968.

30. Cleverdon, C. W., F. W. Lancaster, and J. Mills, "Uncovering Some Facts of Life in Information Retrieval," Special Libraries, Vol. 55, February 1964.

31. ______________ , Report on the Testing and Analysis of an Investigation into the Comparative Efficiency of Indexing Systems, The College of Aeronautics, Cranfield, England, October 1962.

32. ______________ , Factors Determining the Performance of Indexing Systems, Volume 1 and 2, The College of Aeronautics, Cranfield, England, 1966.

33. Cooper, W. S., "A Definition of Relevance for Information Retrieval," Information Storage and Retrieval, Vol. 7, No. 1, June 1971.

34. ____________ , "On Deriving Design Equations for Information Retrieval Systems," Journal of ASIS, November/December 1970.

35. ____________, "Expected Search Length: A Single Measure of Retrieval Effectiveness Based on the Weak Ordering Action of Retrieval Systems," American Documentation, Vol. XIX, January 1968.

36. Cuadra, C., and R. Katter, "Opening the Black Box of Relevance," Journal of Documentation, Vol. 23, 1967.

37. Cuadra, C. A., "*On the Utility of the Relevance Concept*, System Development Corporation, SP-1595, Santa Monica, California, March 1964.

38. Cuadra, C., and R. Katter, et. al., *Experimental Studies of Relevance Judgment*, System Development Corporation, Santa Monica, California, 1967.

39. Curry, H. B., R. Feys, and W. Craig, *Combinatory Logic*, Vol. 1, North Holland Publishing Company, Amsterdam, 1968.

40. DeLuca, A., and S. Termini, "A Definition of a Nonprobabilistic Entropy in the Setting of Fuzzy Sets Theory," *Information and Control*, Vol. 20, 1972.

41. De Solla Price, D. J., "Nations Can Publish or Perish," *Science and Technology*, October 1967.

42. ____________________, *Little Science, Big Science*, Columbia University Press, New York, 1963.

43. ____________________, *Science Since Babylon*, Yale University Press, New Haven, Connecticut, 1961.

44. Doyle, L. B., "Indexing and Abstracting by Association," *American Documentation*, Vol. 13, 1962.

45. ____________ , *Is Relevancy an Adequate Criterion in Retrieval System Evaluation*, System Development Corporation, SP-1262, Santa Monica, California, July 1963.

46. Dym, E. D., "Relevance Predictability: Investigation into Background and Procedure," *Electronic Handling of Information*, Kent and Taulbee (eds.), Academic Press, London, England, 1967.

47. Fairthorne, R. A., "Basic Parameters of Retrieval Tests," *American Documentation Institute Parameters of Information Science*, Spartan Books, Washington, D. C., 1964.

48. __________________, *Towards Information Retrieval*, Butterworths, London, England, 1961.

49. __________________, "Progress in Documentation," *Journal of Documentation*, Vol. 25, No. 4, December 1969.

50. Feller, W., *Introduction to Probability Theory and Its Application*, Vols. 1 and 2, John Wiley, New York, 1950.

51. Fisher, G. H., *Cost Considerations in Systems Analysis*, American Elsevier, New York, 1971.

52. Foskett, A. C., *The Subject Approach to Information*, Archon Books, Handen, Connecticut, 1969.

53. Gazale, M. J., "Irredundant Disjunctive and Conjunctive Forms of a Boolean Function," *IBM Journal of Research and Development*, Vol. 1, 1957.

54. Guiliano, V., "The Interpretation of Word Association," *Symposium on Statistical Association Methods for Mechanized Documentation*, March 17-19, 1964.

55. Giuliano, V. E., and P. E. Jones, Study and Test of a Methodology for Laboratory Evaluation of Message Retrieval Systems, CFSTI, 1966.

56. Goffman, W., and V. Newill, "Methodology for Test and Evaluation of Information Retrieval Systems," Information Storage and Retrieval, Vol. 3, 1966.

57. Goffman, W., and K. Warren, "Dispersion of Papers Among Journals Based on a Mathematical Analysis of Two Diverse Medical Literatures," Nature, March 29, 1969.

58. Good, I. J., "The Decision Theory Approach to the Evaluation of Information Retrieval Systems," Information Storage and Retrieval, Vol. 3, August 1966.

59. Gottschalk, C. F., and D. Desmond, "Worldwide Census of Scientific and Technical Serials," American Documentation, July 1963.

60. Groos, O. V., "Bradford's Law and the Keenan-Atherton Data," Journal of American Documentation, Vol. 19, No. 1, 1967.

61. Gull, C. D., "Seven Years of Work on the Organization of Materials in the Special Library," American Documentation, Vol. 7, October 1956.

62. Harlow, J., and P. Abrahams, An Investigation of the Techniques and Concepts of Information Retrieval, Final Report, Signal Corps, July 31, 1964.

63. Hayes, R. M., "Mathematical Models for Information Retrieval," Natural Language and the Computer, Garvin (ed.), McGraw Hill, New York, 1963.

64. Herdan, G., The Advanced Theory of Language as Choice and Chance, Springer-Verlag, New York, 1966.

65. ________ , Quantitative Linguistics, Butterworths, London, England, 1964.

66. Hertz, D. B., Research Study of Criteria and Procedures for Evaluating Scientific Information Retrieval Systems, Authea Anderson and Co., March 1962.

67. Holt, C., and W. Schrank, "Growth of the Professional Literature in Economics and Other Fields, and Some Implications," American Documentation, January 1968.

68. Houston, N., and E. Wall, "The Distribution of Term Usage in Manipulative Indexes," American Documentation, April 1964.

69. IFD, The ASLIB Cranfield Research Project Report A 7&8, 1968.

70. Iker, H. P., "Solution of Boolean Equations Through the Use of Term Weights to the Base Two," American Documentation, January 1967.

71. Jahoda, G., Information Storage and Retrieval Systems for Individual Researchers, Wiley-Interscience, New York, 1970.

72. Jones, P. E., and R. M. Curtice, "A Framework for Comparing Term Association Measures, American Documentation, July 1963.

73. Jones, P. E., et. al., Papers on Automatic Language Processing: Selected Collection Statistics and Data Analysis, A. D. Little, February 1967.

74. Katter, R. V., "Design and Evaluation of Information Systems," Annual Review of Information Science and Technology, Vol. 4, C. Cuadra (ed.), Encyclopedia Britannica, Inc., Chicago, 1969.

75. Kendall, M. G., "The Bibliography of Operations Research," Operational Research Quarterly, Vol. 11, No. 1/2.

76. Kessler, M. M., Technical Information Flow Patterns, Lincoln Laboratories, MIT.

77. King, D. W., Evaluation During File Development of the Glass Technology Coordinate Index, U.S. Department of Commerce, Patent Office, November 1967.

78. Kochen, M., "System Technology for Information Retrieval," The Growth of Knowledge: Readings on Organization and Retrieval of Information, John Wiley & Sons, New York, 1967.

79. Krauze, T. K., and C. Hillinger, "Citations, References and the Growth of Scientific Literature: A Model of Dynamic Interaction," Journal of ASIS, September/October 1971.

80. Krevitt, B., and B. C. Griffith, "A Comparison of Several Zipf-Type Distributions in Their Goodness of Fit to Language Data," Journal of ASIS, May-June 1972.

81. Kuhns, I. L., "The Continuum of Coefficients of Association," Statistical Association Methods for Mechanized Documentation, Symposium Proceedings, Spartan Books, Washington, D. C., 1964.

82. Lancaster, F. W., Evaluation of the Medlars Demand Search Service, National Library of Medicine, 1968.

83. ______________, Information Retrieval Systems: Characteristics, Testing and Evaluation, John Wiley, New York, 1968.

84. Lancaster, F. W., and W. D. Climenson, "Evaluating the Econometric Efficiency of a Document Retrieval System," Journal of Documentation, Vol. 24, No. 1, 1968.

85. ______________, "The Cost-Effectiveness Analysis of Information Retrieval and Dissemination Systems," Journal of ASIS, January-February 1971.

86. Lesk, M., and G. Salton, "Interactive Search and Retrieval Methods Using Automatic Information Displays," Proceedings of the Spring Joint Computer Conference, 1969.

87. ________, "Word Association in Document Retrieval Systems," American Documentation, January 1969.

88. Leimkuhler, F. F.,"The Bradford Distribution," Journal of Documentation, Vol. 23, No. 3, September 1967.

89. Line, M. B., "The Half-Life of Periodical Literature: Apparent and Real Obsolescence," Journal of Documentation, Vol. 26, March 1970.

90. Litofsky, B., Utility of Automatic Classification Systems for Information Storage and Retrieval, Ph.D. Thesis, University of Pennsylvania, May 1969.

91. Long, J. M., et. al., "Dictionary Buildup and Stability of Word Frequency in a Specialized Medical Area," American Documentation, January 1967.

92. Lowe, T. C., Design Principles for an On-Line Information Retrieval System, Moore School Report No. 67-14, The Moore School of Electrical Engineering, Philadelphia, Pennsylvania, 1966.

93. Mcluhan, M., The Gutenberg Galaxy, University of Toronto Press, Canada, 1962.

94. Mandelbrot, B., "An Informational Theory of the Statistical Structure of Language," Communication Theory, Jackson Willis (ed.), Academia Press, New York, 1953.

95. ____________ , "Simple Games of Strategy Occurring in Communication Through Natural Languages," Transactions of the I.R.E., Professional Group on Information Theory, Vol. 3, March 1954.

96. ____________ , "On the Theory of Word Frequencies and on Related Markovian Models of Discourse," Proceedings of the Twelfth Symposium in Applied Mathematics, R. Jakobson (ed.), American Mathematical Society, Rhode Island, 1961.

97. Maron, M. E., and J. L. Kuhns, "On Relevance, Probabilistic Indexing and Information Retrieval," Journal of ACM, Vol. 7, No. 3, 1960.

98. Maron, M. E., A. Humphry, and J. Meredith, An Information Processing Laboratory for Education and Research in Library Science: Phase I, Institute of Library Research, University of California, Berkeley, June 1969.

99. Marron, H., "On Costing Information Services," Proceedings of the American Society for Information Science, 1969.

100. Martyn, J., and B. C. Vickery, "The Complexity of the Modelling of Information Systems," Journal of Documentation, Vol. 26, No. 3, September 1970.

101. Meadow, C. T., The Analysis of Information Systems--A Programmer's Introduction to Information Retrieval, John Wiley, New York, 1967.

102. Mood, A. M., and F. A. Graybill, Introduction to the Theory of Statistics, McGraw Hill Book Company, New York, 1963.

103. Mott, T. H., Jr., "Determination of the Irredundant Normal Forms of a Truth Function by Iterated Consensus of the Prime Implicants," I.R.E. Transactions on Electronic Computers, Vol. EC-9, 1960.

104. Oettinger, A. G., "An Essay in Information Retrieval or the Birth of a Myth," Information and Control, Vol. 8, 1965.

105. Overmever, L., "An Analysis of Output Costs and Procedures for an Operational Searching System," American Documentation, Vol. 14, 1963.

106. Perry, J. W., A. Kent, and M. M. Berry, Machine Literature Searching, Interscience, New York, 1956.

107. "Proceedings of the American Documentation Institute," Annual Meeting, Vol. 4, New York, October 22-27, 1967.

108. Quine, W. V., "A Way to Simplify Truth Functions," American Mathematics Monthly, Vol. 62, 1955.

109. ________, "The Problem of Simplifying Truth Functions," American Mathematics Monthly, Vol. 59, 1952.

110. Ranganathan, S. R., The Colan Classification, Vol. IV, Rutgers Series on Systems for the Intellectual Organization of Information, S. Artandi (ed.), Rutgers University, 1965.

111. Raver, N., "Performance of Information Retrieval Systems," Information Retrieval, George Schecter (ed.), Thompson Book Company, Washington, D. C., 1967.

112. Rees, A. M., "Evaluation of Information Systems and Services," Annual Review of Information Science and Technology, Vol. 2, C. Caudra (ed.), Interscience, New York, 1968.

113. ________, The Evaluation of Retrieval Systems, Comparative Systems Laboratory, TR-5, Western Reserve University, Cleveland, Ohio, July 1965.

114. Robertson, S. E., "The Parametric Description of Retrieval Tests, Part II," Journal of Documentation, Vol. 23, No. 2, June 1969.

115. Rosenberg, V., "A Study of Statistical Measures for Predicting Index Term Usage," Journal of ASIS, February 1971.

116. Ruspini, E. H., "A New Approach to Clustering," Information and Control, Vol. 15, 1969.

117. Salton, G., Automatic Information Organization and Retrieval, McGraw Hill, New York, 1968.

118. ________, "The Evaluation of Automatic Retrieval Procedures--Selected Results Using the SMART System," American Documentation, Vol. 16, No. 3.

119. ________, The SMART Project--Status Report and Plans: Reports on Evaluation, Clusters, and Feedback, Scientific Report No. ISR-12, Cornell University, New York, 1967.

120. Saracevic, T., "Selected Results from an Inquiry into Testing of Information Retrieval Systems," Journal of ASIS, March-April 1971.

121. Schultz, C., C. Schwartz, and L. Steinberg, "A Comparison of Dictionary Use Within Two Information Retrieval Systems," American Documentation, October 1961.

122. Schultz, L. (ed.), The Information Bazaar; Sixth Annual National Colloquium of Information Retrieval, Mechanical Documentation Service, 1969.

123. Sharp, J. R., Some Fundamentals of Information Retrieval, London House & Maxwell, New York, 1965.

124. Shumway, R. H., "Some Estimation Problems Associated with Evaluating Information Retrieval Systems," Evaluation of Document Retrieval Systems, E. C. Bryant (ed.), Westat Research, Inc., December 31, 1968.

125. Soergel, D., "Mathematical Analysis of Documentation Systems," Information Storage and Retrieval Systems, Vol. 3, 1967.

126. Snyder, M. B., et. al., Methodology for Test and Evaluation of Document Retrieval Systems, Human Sciences Research, Inc., Maclean, Virginia, January 1966.

127. Svenonius, E., "An Experiment in Index Term Frequency," Journal of ASIS, March-April 1972.

128. Swanson, D. R., "On Indexing Depth and Retrieval Effectiveness," Information System Sciences: Proceedings of the Second Congress, Hot Springs, Virginia, November 22-25, 1964, J. Spiegal and D. E. Walker (eds.), Spartan Books, Washington, D. C., 1965.

129. ____________, "Searching Natural Language Text by Computer," Science, Vol. 132, October 21, 1960.

130. ____________, "The Evidence Underlying the Cranfield Results," The Library Quarterly, Vol. XXXV, 1965.

131. Swets, J. A., "Information Retrieval Systems," Science, Vol. 141, July 19, 1963.

132. ___________, Effectiveness of Information Retrieval Methods, Bolt, Bernack, and Newman, June 1967.

133. Switzer, P., "Vector Images in Document Retrieval," Symposium on Statistical Association Methods for Mechanical Documentation, March 17-19, 1964.

134. Szasz, G., Introduction to Lattice Theory, Academic Press, New York, 1965.

135. Taube, M., et. al., Studies in Coordinate Indexing, Vols. 1-4, Documentation, Inc., Bethesda, Maryland, 1953-57.

136. ________, and L. Heilprin, The Relation of the Size of the Question to the Work Accomplished by a Storage and Retrieval System, Documentation, Inc., Washington, D. C., August 1957.

137. Tell, B. V., "Auditing Procedures for Information Retrieval Systems," Proceedings of the 1965 Congress International Federation for Documentation, 31st Annual Meeting: October 7-16, 1965, Spartan Books, Washington, D. C., 1966.

138. Thorne, R. G., "The Efficiency of Subject Catalogues and the Cost of Information Searches," Journal of Documentation, Vol. 11, September 1955.

139. Tinker, J. F., "Imprecision in Meaning Measured by Inconsistency in Indexing," American Documentation, April 1966.

140. Vickery, B. C., On Retrieval System Theory, Butterworths, London, England, 1961.

141. ______________, "Bradford's Law of Scattering," Journal of Documentation, Vol. 4, No. 3, 1948.

142. Voigt, M. J., "The Researcher and His Sources of Scientific Information," Libri, Vol. 9, 1959.

143. Wall, E., "Further Implications of the Distribution of Index Term Usage," Proceedings of the American Documentation Institute, Vol. 1, Parameters of Information Science Annual Meeting," Philadelphia, Pennsylvania, October 5-8, 1964.

144. ______________, "Indexing Control," TICA Conference, Drexel Institute of Technology, June 15-17, 1964, Spartan Books, Washington, D. C., 1964.

145. Weaver, W., "Recent Contributions to the Mathematical Theory of Communication," The Mathematical Theory of Communications, C. E. Shannan and W. Weaver (eds.), University of Illinois Press, Urbana, Illinois, 1949.

146. Webster, R., "A Note on Dictionary Searching," Information Storage and Retrieval, Vol. 5, 1969.

147. Westat Research, Inc., Procedural Guide for the Evaluation of Document Retrieval Systems, December 1968.

148. Wiederkehr, R. R., "Search Characteristic Curves," Evaluation of Document Retrieval Systems, E. Bryant (ed.), Westat Research, Inc., December 31, 1968.

149. Wiener, N., The Human Use of Human Beings; Cybernetics and Society, Houghton Mifflin, Boston, 1950.

150. Yule, G. H., "On Measuring Association Between Attributes," Journal of the Royal Statistical Society, Vol. 75, 1912.

151. Zadeh, L. A., Fuzzy Sets, Electronics Research Laboratory, Report No. 64-44, University of California, Berkeley, November 1964.

152. Zunde, P., and V. Slamecka, "Distribution of Indexing Terms for Maximum Efficiency of Information Transmission," American Documentation, April 1967.

153. Zipf, G. K., Human Behavior and the Principle of Least Effort, Addison Wesly, Cambridge, Massachusetts, 1949.

Appendix A

GLOSSARY

GLOSSARY

Boolean Algebra -- a Boolean Algebra is defined as a distributive lattice in which each element "a" has a complement defined by its negation. This structure, for a defined set T and its elements (A,B,...), is defined in terms of the following operations. Conjunction; $C = A \cdot B$, the subset of subclass of all index terms or elements of T that are both in the subsets of A and B. Disjunction; $D = A+B$, the subset of all index terms or elements of T which are either in subset A or subset B. Negation; $N = -B$ or B, the subset of all index terms in T which are not in subset B.

Bradfords Law of Literary Yield or Scatter -- if periodicals are ranked into N groups, each yielding the same number of articles as a specified topic, the number of periodicals in each group will increase geometrically, as per: $1:n:n^2$.

Coordinate Index -- an index system in which the descriptor terms are manipulated. There are two classes of coordinate index systems:

a) Pre-coordinate -- those DRSs in which the coordination of the descriptors takes place during the inquiry generation process.

b) Post-coordinate -- those DRSs in which the coordination of the descriptors takes place during the inquiry generation process.

Document -- any discrete unit of information -- articles, reports, recordings, etc.

Document Retrieval Systems -- a class of information retrieval systems solely concerned with the subject analysis of document content, the storage of a set of official surrogates "defining" document content, and the "mechanical" search of the surrogate set to identify or select those documents most "relevant" to a user's formal request.

Facet Index -- a composite index of an item by combining in a prescribed manner the terms derived from separate relational indexing examinations.

Indexing -- the process in which documents are analyzed, and terms indicating subject content are assigned or derived.

Mandelbrot-Estoup-Zipff Relationship -- the term frequency of use f(r) versus rank (r) distribution in a language is a decreasing convex function in log-log space, and is of the form:

$$f(r) = K(r+B)^{-\alpha}$$

Uniterms, Keywords, Descriptors -- words or word-pairs extracted from a document that are used to identify the subject content of the document.

Word Relationships -- there are four operational word relationship categories that can be employed in DRSs.

(1) Semantic relationships which manifest the meaning and context of terms within a language,

(2) Syntatic relationships which arise from terms as members of word classes and with the class relationships in a structural (grammatical) sense,

(3) Syndetic relationships which measure the manner by which words that are conjunctively coordinated with a given or base term cross-reference one another, and

(4) Statistical relationships which measure the frequency of occurrence of terms in a document.

Zipf "Law" of Term Usage -- the relationship between the frequency of use $f(r)$ of a term and its rank (r) in a language based on Zipf's Principle of Least Effort, and is of the form:

$$f(r) = Kr^{-1}$$

Appendix B

INSTITUTE OF LIBRARY RESEARCH -- TEST SYSTEM CHARACTERISTICS

- o Thesaurus Listing (Sample)
- o Document Descriptor Listing (Sample)

THESAURUS LISTING SAMPLE

SUBJECT AUTHORITY LIST (98)

ABBREVIATIONS

S = SEE
SA = SEE ALSO
SN = IN THE SENSE OF (I.E. SCOPE NOTE)
* = NO DOCUMENTS YET INDEXED WITH THIS TERM
+ = TERM NOT ALLOWED, RELATED TERM TO BE USED

*ABBREVIATION
ABSTRACT
ABSTRACTING
ACCESS
ACCESSION NUMBER
ACCURACY
ACQUISITION
ADDRESS
ADMINISTRATION
ALGEBRA
+ALGOL
    S PROG. LANGUAGE
ALGORITHM
ALPHABETIC
ALPHABETIC ORDER
ALPHANUMERIC
*ALTERNATIVES
AMBIGUITY
ANALOGY
ANALYSIS
ANSWER
*ANTHOLOGY
    SA BIBLIOGRAPHY
APPLICATION
+ARITHMETIC
    S MATHEMATICS
ARRAY
+ARTICLE
    S DOCUMENT
ARTIFICIAL INTEL
ASSIGNED
ASSOCIATION
ASSOCIATIVE
+ATTRIBUTE
    S CHARACTERISTIC
AUTHOR
AUTHORITY LIST
    SA THESAURUS
AUTO ABSTRACTING
AUTO. INDEXING
AUTOMATIC
AUTOMATION
    SA MECHANIZATION

BATCH PROCESSING
BIBLIOGRAPHIC
BIBLIOGRAPHY
    SA ANTHOLOGY
BINARY
BOOK
BOOLEAN
    SA LOGICAL

CALL NUMBER
CANONICAL
    SA NORMALIZED
CARD
CARD CATALOG
CATALOG
CATALOGING
CATEGORIES
CENTERS
CENTRALIZED
CHARACTERISTIC

- CHEMICAL
- CIRCULATION
- CITATION
- CITATION INDEX
- *CLAIM
  - SA COPYRIGHT
  - SA PATENT
- CLASSIF. SCHEME
- CLASSIFICATION
- CLERICAL
- +CLUE WORD
  - S KEYWORD
- CLUMP
- CLUSTER
- CO-OCCURRENCE
- +COBOL
  - S PROG. LANGUAGE
- CODE
  - SN MEDIA DESIGNATION
- CODING
  - SN COMPUTER CODING
- COEFFECIENT
- COLLECTION
- *COLLOQUIUM
  - SA CONFERENCE
  - SA MEETING
  - SA SYMPOSIUM
- COMBINATIONS
- +COMIT
  - S PROG. LANGUAGE
- COMMUNICATION
- COMP LINGUISTICS
- COMPARISON
- COMPUTER
- CONCEPT
- CONCORDANCE
- CONDITIONAL PROB
- CONFERENCE
  - SA COLLOQUIUM
  - SA MEETING
  - SA SYMPOSIUM
- CONNECTION
- +CONSECUTIVE
  - S ORDER
- +CONSOLE
  - S REMOTE TERMINAL
- CONTENT
- CONTENT ANALYSIS
- CONTEXT
- CONTROL
- CONTROLLED
- CONVENTIONAL
- CONVERSION
- COORDINATE
- COORDINATE INDEX
  - SA UNITERM SYSTEM
- *COPYRIGHT
  - SA CLAIM
  - SA PATENT
- +CORE
  - S STORAGE
- CORRELATION
- COST
- COUNT
- COUPLING
- CRANFIELD
- CRITERIA
- CRITICAL
  - SN REVIEWING, NOT VITAL
- CROSS REFERENCE
- CURRENT AWARENES
- CURRICULUM
- +CUSTOMER
  - S USER

- DATA
- *DECENTRALIZATION
- DECISION THEORY
- DEDUCTIVE
- DEGREE
- DEPTH OF INDEXIN
- DESCRIPTIVE
- DESCRIPTOR
  - SA KEYWORD
  - SA TAG
  - SA TERM
- DESIGN
  - SA PLANNING
- DICTIONARY
- +DIFFERENCE
  - S COMPARISON
- +DIGITAL COMPUTER
  - S COMPUTER
- DISCRIMINANT
- +DISPLAY
  - S REMOTE TERMINAL
- DISSEMINATION
- *DISSERTATION
- DOCUMENT
  - SA JOURNAL
- DOCUMENTATION
- DUAL DICTIONARY

- +ECONOMICS
  - S COST
- EDITING
- EDUCATION
- EFFECTIVENESS
  - SA EFFICIENCY

```
 EFFICIENCY
      SA  EFFECTIVENESS
+ELECTRONIC COMPUTER
      S   COMPUTER
+EMPIRICAL
      S   EXPERIMENT
+ENCODING
      S   CODING
 ENTROPY
 ENTRY
      SN  ACCESS POINT
 ERROR
 EVALUATION
      SA  TEST
      SA  UTILITY
      SA  VALUE
 EXPERIMENT
 EXTRACT


 FACET
 FACETED CLASSIF.
 FACT RETRIEVAL
+FACTOR ANALYSIS
      S   STAT. METHOD
 FALSE DROP
 FEEDBACK
 FILE
      SA  LIST
      SA  STRING
 FILE ORGANIZATIO
 FLOW OF INFO.
 FORMAT
+FORTRAN
      S   PROG. LANGUAGE
 FREQUENCY
 FUNCTION
      SN  OPERATIONAL, NOT
          MATHMATICAL


 GENERAL
 GENERATION
      SN  PRODUCTION
 GENERIC
+GOAL
      S   OBJECTIVE
 GOVERNMENT
 GRAMMAR
 GRAPH
      SN  MATHEMATICAL GRAPH
      SA  TABLE
 GRAPHICS
      SN  GRAPHIC MATERIALS E.G.
          PHOTOS.
```

```
+GROUP
      S   CLUMP


 HARDWARE
      SN  COMPUTERS, MICROFILM
          EQUIPMENT, ETC.
      SA  MECHANICAL
+HEADINGS
      S   SUBJECT HEADING
 HIERARCHY
 HISTORICAL
+HUMAN
      S   MANUAL
+HUMAN INDEXING
      S   MANUAL INDEXING


*IDENTICAL
 IDENTIFICATION
 ILLUSTRATION
*IMPLEMENTATION
 INDEPENDENT
 INDEX
 INDEXING
 INFERENCE
 INFO. RETRIEVAL
 INFO. SCIENCE
 INFORMATION
 INPUT
+INQUIRER
      S   USER
+INQUIRY
      S   QUESTION
+INSTRUCTION
      S   EDUCATION
 INTELLECTUAL
 INTERDISCIPLINAR
 INTERFACE
 INTERPRET
+INTERROGATE
      S   QUESTION
+INTERSECTION
      S   VENN DIAGRAM
 INTRODUCTORY
 INTUITIVE
 INVENTORY
*INVERTED
 IRRELEVANT
+ITEM
      S   DOCUMENT
 ITERATIVE

      SA  RECURSIVE
```

- JOURNAL
  - SA DOCUMENT

- KEYPUNCH
- KEYWORD
  - SA DESCRIPTOR
  - SA TAG
  - SA TERM
- KWIC

- LANGUAGE
- LARGE
- LATTICE
- LAW
- +LEVEL
  - S DEGREE
- +LEXICAL
  - S ALPHABETIC
- +LEXICON
  - S DICTIONARY
- LIBRARIAN
- LIBRARY
- LINGUISTIC
- LINK
- LIST
  - SA FILE
  - SA STRING
- LITERATURE
- LOGIC
- LOGICAL
  - SA BOOLEAN

- +MACHINE
  - S HARDWARE
- MACHINE-READABLE
- +MAGNETIC TAPE
  - S STORAGE
- MAN-MACHINE
- MANUAL
- MANUAL INDEXING
- MATCH
- MATHEMATICAL
- MATHEMATICS
  - SA PROBABILITY
- MATRIX
- MEANING
- MEASURE
- MECHANICAL
  - SA HARDWARE
- MECHANIZATION
  - SA AUTOMATION
- MEDIUM
- MEETING
  - SA COLLOQUIUM
  - SA CONFERENCE
  - SA SYMPOSIUM
- +MEMORY
  - S STORAGE
- METHODOLOGY
- +METRIC
  - S MEASURE
- MICROFICHE
- MICROFILM
- MODEL
  - SA SIMULATION
- MODIFICATION
- MULTIPLE

- NATIONAL
- NATURAL
- NATURAL LANGUAGE
- NEEDS
- NETWORK
  - SN ORGANIZATIONAL STRUCTURE
  - SA ORGANIZATION
- NOISE
- +NOMENCLATURE
  - S NOTATION
- NON-CONVENTIONAL
- NON-DISCRIMINANT
- NON-FILE
- NON-RANDOM
- NON-RELEVENT
- *NORMALIZED
  - SA CANONICAL
- NOTATION
  - SA TERMINOLOGY
- NUMBER
- NUMERIC

- OBJECTIVE
  - SN GOAL, NOT AS OPPOSED
    TO SUBJECTIVE
- *OCCURRENCE
- OFF-LINE
- ON-LINE
- OPERATION
- OPTIMIZATION
- ORDER
- ORGANIZATION
  - SA NETWORK
- OUTPUT

- +PAIR
  - S WORD ASSOCIATION

```
+PAPER
      S    DOCUMENT
 PARAMETER
      SA   VARIABLE
 PARSE
 PATENT
      SA   CLAIM
      SA   COPYRIGHT
 PATTERN
 PERFORMANCE
+PERIODICAL
      S    JOURNAL
 PERMUTED
 PERTINENT
      SA   RELEVANT
 PHILOSOPHY
      SA   POLICY
+PHOTO
      S    GRAPHICS
 PLANNING
      SA   DESIGN
+PLOT
      S    GRAPH
+POLICY
      SA   PHILOSOPHY
+POPULATION
      S    COLLECTION
 PRECISION
 PREDICTION
*PRINCIPLE
+PRINT-OUT
      S    OUTPUT
 PRINTING
+PRIVACY
      S    SECRECY
 PROBABILITY
      SA   MATHEMATICS
 PROCEDURE
 PROCEEDINGS
 PROCESSING
 PROFILE
 PROG. LANGUAGE
 PROGRAM
      SN   COMPUTER PROGRAM
      SA   ROUTINE
      SA   SOFTWARE
      SA   SUBROUTINE
 PROGRAMMED
+PROPERTY
      S    CHARACTERISTIC
 PSYCHOLOGY
+PUBLICATION
      S    DOCUMENT
 PUNCHED
+PUNCHED-CARD
      S    STORAGE
```

```
 PUNCTUATION
+PURPOSE
      S    OBJECTIVE


 QUALITATIVE
      SA   SUBJECTIVE
 QUANTITATIVE
+QUERY
      S    QUESTION
 QUESTION
      SN   BOTH NOUN AND VERB
 QUESTION-ANSWER


 RANDOM
 RANDOM-ACCESS
 RANK
 READING
 REAL-TIME
 RECALL
 RECOGNITION
 RECORD
+RECORDED INFO.
      S    RECORD
 RECURSIVE
      SA   ITERATIVE
 REDUNDANCY
 REFERENCE
*REJECTION
 RELATED
 RELATIONSHIP
 RELATIVE
 RELEVANCE
 RELEVANT
      SA   PERTINENT
+REMOTE TELETYPES
      S    REMOTE TERMINAL
 REMOTE TERMINAL
      SA   VISUAL DIS. CON.
+REPORT
      S    DOCUMENT
+REQUEST
      S    QUESTION
 RESEARCH
+RESPONSE
      S    ANSWER
 RESPONSE TIME
 RETRIEVAL
 RETRIEVAL SYSTEM
 REVIEW
      SA   SUMMARY
      SA   SURVEY
 ROLE
 ROUTINE
      SN  COMPUTER ROUTINE
      SA  PROGRAM
      SA  SOFTWARE
      SA  SUBROUTINE
 RULE


 SAMPLE
 SCANNING
 SCIENTIFIC
 SCOPE NOTE
 SEARCH CRITERIA
 SEARCH STRATEGY
 SEARCHING
*SECRECY
 SEE ALSO
      SN  AS USED IN CATALOGING
 SEE-REFERENCE
 SELECTION
 SELECTIVE DISSEM
 SEMANTIC
      SA  SYNTAX
 SEQUENCE
+SERIAL
      S   JOURNAL
 SERVICE
 SET THEORY
 SETS
 SHELFLIST
 SIGNIFICANCE
 SIMULATION
      SA  MODEL
 SIZE
 SMALL
 SOCIAL IMPLIC.
 SOFTWARE
      SA  PROGRAM
      SA  ROUTINE
      SA  SUBROUTINE
 SORTING
 SOURCE
 SPECIALIZED
 SPECIFICITY
 STANDARDIZATION
 STAT ASSOCIATION
 STAT. ANALYSIS
      SA  STAT. METHOD
 STAT. METHOD
      SA  STAT. ANALYSIS
 STATE-OF-THE-ART
 STATISTICAL
+STOCHASTIC
      S   RANDOM
 STORAGE
```

```
 STRING
      SA  FILE
      SA  LIST
 STRUCTURE
 SUBJECT
 SUBJECT HEADING
 SUBJECT INDEXING
 SUBJECT-CATALOG.
+SUBJECTIVE
      SA  QUALITATIVE
 SUBROUTINE
      SA  PROGRAM
      SA  ROUTINE
      SA  SOFTWARE
 SUMMARY
      SA  REVIEW
      SA  SURVEY
 SURVEY
      SA  REVIEW
      SA  SUMMARY
 SYMBOL
 SYMBOLIC LOGIC
 SYMPOSIUM
      SA   COLLOQUIUM
      SA   CONFERENCE
      SA   MEETING
 SYNONYM
 SYNTACTIC ANAL.
 SYNTAX
      SA  SEMANTIC
 SYSTEM


 TABLE
      SA  GRAPH
 TAG
      SA  DESCRIPTOR
      SA  KEYWORD
      SA  TERM
+TAPE
      S   STORAGE
+TEACHING
      S   EDUCATION
 TECHNICAL
 TECHNICAL REPORT
 TECHNOLOGY
 TELEGRAPHIC ABS.
 TERM
      SA  DESCRIPTOR
      SA  KEYWORD
      SA  TAG
+TERMINAL
      S   REMOTE TERMINAL
 TERMINOLOGY
      SA  NOTATION
 TEST
      SA  EVALUATION
      SA  UTILITY
      SA  VALUE
 TEXT
 THEORY
 THESAURUS
      SA  AUTHORITY LIST
 TIME
 TIME-SHARING
 TITLE
*TOPIC
      S    SUBJECT
 TRANSFORMATION
 TRANSLATION
*TRANSLITERATION
 TRANSMISSION
 TREE
 TREE STRUCTURE
 TRUNCATION
*TYPE STYLE
 TYPE-SETTING
*TYPOGRAPHICAL


*UNION
      SN  SET THEORY UNION
      S   VENN DIAGRAM
*UNION CATALOG
*UNITERM
      S    DESCRIPTOR
 UNITERM SYSTEM
      SA  COORDINATE INDEX
 UPDATING
 USER
 UTILITY
      SA  EVALUATION
      SA  TEST
      SA  VALUE


 VALIDATION
 VALUE
      SA  EVAUATION
      SA  TEST
      SA  UTILITY
 VARIABLE
      SA  PARAMETER
 VECTOR
 VENN DIAGRAM
*VISUAL DIS. CON.
      SA  REMOTE TERMINAL
 VOCABULARY


 WEIGHT
 WEIGHT INDEXING
 WORD
 WORD ASSOCIATION
 WORD FREQUENCY
*WORD PAIRS
      S    WORD ASSOCIATION
```

DOCUMENT DESCRIPTOR LISTING (98)

Document I.D.

DESCRIPTORS

```
A013101LDACCESS             ALGOPITHM           ASSIGNED            COST
A013102LDDATA               FILE                INFORMATION         LANGUAGE
A013103LDLIST               NOTATION            OPERATION           PROCEDURE
A013104LDPROG. LANGUAGE     PROGRAM             SETS                STORAGE
A013105LDSTRING             STRUCTURE           SYNTAX              SYSTEM
A013106LDVARIABLE                                                              *

A013201LDACCFSS             ALGORITHM           COMMUNICATION       COMPUTER
A013202LDCONTEXT            DATA                FLOW OF INFO.       GENERATION
A013203LDGRAMMAR            INFO. RETRIEVAL     INFORMATION         INTERPRET
A013204LDNATURAL LANGUAGE   OUTPUT              PARSE               QUESTION-ANSWER
A013205LDRFLEVANT           SEMANTIC            STORAGE             SURVEY
A013206LDSYNTACTIC ANAL.    SYNTAX              SYSTEM              TIME-SHARING
A013207LDTRANSFCRMATION                                                        *

A013301LDABSTRACTING        ALGORITHM           ANALYSIS            COMP LINGUISTICS
A013302LDCONFFRENCE         EDITING             EVALUATION          INFO. RETRIEVAL
A013303LDLINGUISTIC         LOGIC               MATCH               NATURAL LANGUAGE
A013304LDPARSE              PROG. LANGUAGE      PROGRAM             QUESTION-ANSWER
A013305LDSYMBOLIC LOGIC     TFCHNICAL           TIME-SHARING        TRANSLATION    *

A013401LDALGORITHM          COMPUTER            CONFERENCE          ERROR
A013402LDINTERPRET          MAN-MACHINE         MATHEMATICAL        NATURAL LANGUAGE
A013403LDNOISF              NOTATION            PROG. LANGUAGE      PROGRAM
A013404LDREDUNDANCY         SEMANTIC            SOFTWARE            SYNTAX
A013405LDSYSTFM             TRANSLATION         USER                WORD           *

A013501LDACCESS             BIBLIOGRAPHY        CENTERS             CIRCULATION
A013502LDDOCUMENT           FLOW OF INFO.       GENERAL             INFO. RETRIEVAL
A013503LDLIBRARIAN          LIBRARY             MECHANIZATION       REMOTE TERMINAL
A013504LDRESEARCH           SCIENTIFIC          SEARCHING           SERVICE
A013505LDTECHNOLOGY         TRANSMISSION                                       *

A013601LDACQUISITION        ANALYSIS            CIRCULATION         COMMUNICATION
A013602LDLIBRARY            MEASURF             MEETING             PATTERN
A013603LDRETRIEVAL          SERVICF             SYSTEM                         *

A013701LDACCESSION NUMBER   BOOK                CLASSIFICATION      LIBRARY
A013702LDRETRIEVAL          SIZE                SUBJECT                        *

B001201LDAUTO ABSTRACTING   BIBLIOGRAPHIC       COMMUNICATION       LANGUAGE
B001202LDLINGUISTIC         NATURAL             STORAGE             SYSTEM
B001203LDTRANSLATION                                                           *

B001301LDABSTRACTING        ASSOCIATION         CLASSIFICATION      DATA
B001302LDDICTIONARY         FREQUENCY           INDEX               INFORMATION
B001303LDLIBRARY            LITERATURE          MICROFILM           NETWORK        *

B001401LDDOCUMENT           INDEXING            INFO. RETRIEVAL     MICROFILM
B001402LDSCANNING           STORAGE             TERM                TRANSLATION    *

B001501LDAUTOMATION         COMMUNICATION       DISSEMINATION       DOCUMENT
B001502LDINFO. RETRIFVAL    INFORMATION         INPUT               OUTPUT
B001503LDQUESTICN           RETRIEVAL           SIGNIFICANCE        THESAURUS      *
```

Appendix C

SAMPLE DATA BASE CHARACTERISTICS

- Term Frequency of Use Listing
- Depth of Indexing Listing
- Term-Document Matrix in Condensed Array Format

TERM FREQUENCY OF USE FOR SAMPLE DATA BASE

Term	Use	Term	Use	Term	Use
		51	4	102	13
1	0	52	3	103	3
2	4	53	0	104	15
3	1	54	6	105	2
4	8	55	20	106	1
5	0	56	1	107	0
6	0	57	5	108	32
7	0	58	7	109	1
8	2	59	8	110	0
9	0	60	8	111	2
10	5	61	5	112	3
11	9	62	9	113	4
12	2	63	5	114	10
13	0	64	0	115	3
14	0	65	1	116	3
15	0	66	10	117	3
16	2	67	1	118	26
17	1	68	4	119	12
18	14	69	27	120	1
19	2	70	7	121	1
20	0	71	0	122	3
21	3	72	0	123	1
22	1	73	1	124	3
23	0	74	1	125	5
24	0	75	5	126	8
25	12	76	4	127	2
26	3	77	3	128	3
27	1	78	2	129	1
28	1	79	3	130	11
29	3	80	1	131	3
30	8	81	0	132	2
31	12	82	4	133	1
32	1	83	10	134	4
33	0	84	0	135	0
34	5	85	6	136	6
35	4	86	6	137	8
36	1	87	1	138	0
37	4	88	1	139	8
38	5	89	3	140	7
39	1	90	3	141	1
40	0	91	1	142	0
41	1	92	2	143	0
42	1	93	2	144	3
43	3	94	2	145	0
44	1	95	5	146	0
45	5	96	1	147	15
46	2	97	2	148	26
47	0	98	0	149	3
48	2	99	6	150	21
49	2	100	1	151	4
50	1	101	1	152	22

Term	Use	Term	Use	Term	Use
153	14	204	4	255	1
154	2	205	3	256	1
155	1	206	0	257	13
156	1	207	1	258	2
157	1	208	1	259	5
158	4	209	1	260	1
159	1	210	1	261	1
160	1	211	0	262	8
161	1	212	4	263	0
162	2	213	1	264	0
163	1	214	1	265	15
164	4	215	1	266	2
165	0	216	0	267	20
166	6	217	1	268	12
167	4	218	3	269	4
168	13	219	3	270	8
169	1	220	1	271	0
170	5	221	5	272	21
171	0	222	1	273	14
172	2	223	8	274	0
173	8	224	0	275	3
174	3	225	2	276	0
175	2	226	0	277	1
176	3	227	2	278	2
177	7	228	7	279	3
178	3	229	3	280	7
179	1	230	1	281	0
180	1	231	1	282	2
181	4	232	0	283	9
182	3	233	10	284	24
183	1	234	5	285	1
184	6	235	0	286	0
185	6	236	0	287	0
186	5	237	9	288	1
187	11	238	4	289	1
188	3	239	4	290	13
189	17	240	8	291	4
190	2	241	2	292	3
191	3	242	3	293	1
192	1	243	8	294	4
193	0	244	0	295	0
194	4	245	1	296	1
195	0	246	2	297	0
196	1	247	0	298	0
197	6	248	1	299	1
198	1	249	3	300	1
199	0	250	17	301	3
200	1	251	5	302	1
201	1	252	4	303	3
202	7	253	1	304	0
203	2	254	6	305	3

Term	Use	Term	Use
306	1	339	12
307	2	340	2
308	0	341	5
309	4	342	6
310	1	343	4
311	16	344	5
312	12	345	0
313	1	346	1
314	12	347	3
315	6	348	3
316	4	349	0
317	6	350	0
318	0	351	0
319	1	352	0
320	0	353	0
321	3	354	6
322	7	355	0
323	1	356	11
324	3	357	4
325	3	358	1
326	6	359	6
327	6	360	2
328	22	361	2
329	2	362	0
330	2	363	0
331	5	364	5
332	1	365	7
333	5	366	4
334	0	367	7
335	1	368	9
336	5	369	3
337	4		
338	9		

DEPTH OF INDEXING DISTRIBUTION

Document	Depth of Indexing	Document	Depth of Indexing
1	27	52	37
2	11	53	12
3	13	54	14
4	14	55	19
5	10	56	3
6	18	57	8
7	12	58	10
8	16	59	12
9	12	60	10
10	15	61	10
11	16	62	16
12	15	63	8
13	26	64	17
14	12	65	12
15	15	66	15
16	12	67	2
17	11	68	15
18	10	69	23
19	11	70	17
20	7	71	12
21	19	72	10
22	16	73	14
23	24	74	9
24	21	75	9
25	15	76	14
26	13	77	10
27	14	78	11
28	13	79	14
29	13	80	15
30	24	81	16
31	18	82	8
32	12	83	11
33	17	84	11
34	18	85	14
35	14	86	15
36	14	87	9
37	25	88	21
38	8	89	14
39	10	90	14
40	12	91	2
41	14	92	14
42	3	93	20
43	8	94	12
44	17	95	19
45	26	96	15
46	13	97	19
47	16	98	12
48	34	99	13
49	22	100	3
50	18	101	16
51	10	102	13

TERM - DOCUMENT MATRIX FOR SAMPLE DATA BASE - IN CONDENSED ARRAY FORM

*Document I.D. Dummy Column
XXX YYY ZZZ Descriptor I.D.

Interpret as document XXX is assigned descriptor ZZZ.

```
*1   1 18  1   2 29  1   3 30  1   4 78  1   5 79  1   6 86  1   7 91  1   8102
 1   9118  1 10120  1 11125  1 12130  1 13148  1 14149  1 15150  1 16170
 1  17177  1 18185  1 19191  1 20262  1 21267  1 22285  1 23290  1 24321
 1  25322  1 26338  1 27343  2   1 34  2   2 57  2   3 68  2   4 88  2   5 90
 2   6108  2   7118  2   8144  2   9248  2 10265  2 11336  3   1 58  3   2 59
 3   3 62  3   4 70  3   5 85  3   6118  3   7119  3   8149  3   9152  3 10185
 3  11186  3 12314  3 13356  4   1 30  4   2 55  4   3 69  4   4 75  4   5115
 4   6119  4   7148  4   8149  4   9166  4 10265  4 11290  4 12311  4 13315
 4  14366  5   1 34  5   2 51  5   3108  5   4130  5   5144  5   6204  5   7227
 5   8265  5   9280  5 10311  6   1 25  6   2 31  6   3104  6   4118  6   5125
 6   6137  6   7140  6   8148  6   9153  6 10189  6 11198  6 12220  6 13233
 6  14257  6 15267  6 16272  6 17284  6 18311  7   1 86  7   2 97  7   3 99
 7   4114  7   5118  7   6124  7   7150  7   8189  7   9267  7 10273  7 11328
 7  12359  8   1 32  8   2 55  8   3094  8   4112  8   5131  8   6168  8   7173
 8   8215  8   9228  8 10231  8 11270  8 12300  8 13311  8 14322  8 15333
 8  16357  9   1 25  9   2 48  9   3 62  9   4 85  9   5 99  9   6189  9   7249
 9   8252  9   9265  9 10311  9 11331  9 12360 10   1 61 10   2 66 10   3 86
10   4115 10   5117 10   6126 10   7152 10   8186 10   9189 10 10205 10 11237
10  12268 10 13254 10 14291 10 15338 11   1 18 11   2 31 11   3 55 11   4 62
11   5 76 11   6 83 11   7105 11   8108 11   9130 11 10131 11 11367 11 12311
11  13356 11 14359 11 15365 11 16367 12   1 21 12   2 31 12   3 55 12   4 57
12   5 58 12   6 59 12   7 62 12   8 95 12   9118 12 10170 12 11187 12 12243
12  13272 12 14338 12 15339 13   1  2 13   2 11 13   3 19 13   4 69 13   5 93
13   6106 13   7108 13   8125 13   9130 13 10137 13 11152 13 12164 13 13177
13  14184 13 15194 13 16205 13 17237 13 18241 13 19273 13 20259 13 21296
13  22329 13 23359 13 24365 13 25366 13 26368 14   1100 14   2102 14   3108
14   4119 14   5137 14   6188 14   7189 14   8233 14   9267 14 10272 14 11280
14  12283 15   1 31 15   2 49 15   3 55 15   4119 15   5126 15   6139 15   7152
15   8176 15   9250 15 10253 15 11256 15 12269 15 13284 15 14328 15 15341
16   1 70 16   2 83 16   3118 16   4119 16   5175 16   6189 16   7233 16   8257
16   9267 16 10272 16 11275 16 12283 17   1 10 17   2 38 17   3 55 17   4108
17   5150 17   6170 17   7185 17   8189 17   9267 17 10272 17 11212 18   1 18
18   2 51 18   3 52 18   4 60 18   5177 18   6230 18   7265 18   8280 18   9342
18  10356 19   1 18 19   2 48 19   3 69 19   4115 19   5152 19   6185 19   7189
19   8197 19   9237 19 10338 19 11339 20   1 31 20   2 49 20   3 69 20   4153
20   5243 20   6283 20   7284 21   1 41 21   2 54 21   3 60 21   4108 21   5109
21   6124 21   7130 21   8150 21   9152 21 10212 21 11221 21 12246 21 13250
21  14252 21 15268 21 16284 21 17291 21 18312 21 19328 22   1 21 22   2 57
22   3 86 22   4114 22   5127 22   6140 22   7150 22   8169 22   9197 22 10219
22  11237 22 12249 22 13270 22 14284 22 15340 22 16347023   1 11 23   2 18
23   3 31 23   4 55 23   5 57 23   6 58 23   7 69 23   8104 23   9108 23 10118
23  11148 23 12152 23 13136 23 14168 23 15202 23 16223 23 17240 23 18242
```

```
23 19267 23 20272 23 21273 23 22310 23 23311 23 24321 32 25326 23 26339
23 27341 23 28344 24  1 26 24  2 82 24  3 86 24  4 95 24  5114 24  6118
24  7119 24  8147 24  9150 24 10152 24 11184 24 12197 24 132 3 24 14228
24 15233 24 16257 24 17283 24 18284 24 19328 24 20356 24 21357 25  1 25
25  2 30 25  3 69 25  4108 25  5111 25  6118 25  7148 25  8153 25  9189
25 10268 25 11272 25 12257 25 13284 25 14328 25 15365 26  1 25 26  2 59
26  3130 26  4137 26  5147 26  6185 26  7189 26  8233 26  9254 26 10268
26 11272 26 12311 26 13339 27  1  4 27  2 79 27  3104 27  4132 27  5136
27  6150 27  7174 27  8202 27  9260 27 10328 27 11338 27 12343 27 13356
27 14357 28  1 10 28  2 28 28  3 66 28  4137 28  5152 28  6186 28  7187
28  8204 28  9265 28 10293 28 11314 28 12331 28 13338 29  1  4 29  2 66
29  3 59 29  4126 29  5158 29  6218 29  7219 29  8243 29  9269 29 10292
29 11328 29 12341 29 13357 30  1  4 30  2 11 30  3 66 30  4 69 30  5 75
30  6128 30  7133 30  8136 30  9150 30 10152 30 11157 30 12202 30 13223
30 14225 30 15251 30 16268 30 17290 30 18312 30 19321 30 20326 30 21327
30 22328 30 23341 30 24343 31  1  4 31  2 35 31  3 46 31  4 50 31  5108
31  6128 31  7132 31  8150 31  9172 31 10173 31 11191 31 12269 31 13270
31 14280 31 15284 31 16292 31 17333 31 18346 32  1  3 32  2 25 32  3 55
32  4 95 32  5104 32  6130 32  7147 32  8152 32  9173 32 10177 32 11196
32 12204 33  1 18 33  2 38 33  3 55 33  4 61 33  5 75 33  6148 33  7152
33  8153 33  9168 33 10223 33 11250 33 12265 33 13275 33 14284 33 15280
33 16314 33 17322 34  1 18 34  2 31 34  3 61 34  4 63 34  5 66 34  6104
34  7108 34  8114 34  9147 34 10151 34 11177 34 12191 34 13238 34 14268
34 15279 34 16284 34 17311 34 18343 35  1 55 35  2 60 35  3 66 35  4108
35  5147 35  6152 35  7153 35  8176 35  9223 35 10272 35 11284 35 12312
35 13315 35 14328 36  1 82 36  2 96 36  3116 36  4134 36  5139 36  6140
36  7147 36  8150 36  9170 36 10187 36 11221 36 12265 36 13267 36 14272
36 16279 36 16290 36 17312 36 18314 36 19328 37  1 12 37  2 25 37  3 35
37  4 37 37  5 60 37  6 63 37  7 80 37  8 83 37  9102 37 10108 37 11148
37 12164 37 13176 37 14177 37 15182 37 16187 37 17188 37 18190 37 19262
37 20272 37 21312 37 22322 37 23329 37 24339 37 25354 38  1  4 38  2 61
38  3102 38  4239 38  5272 38  6284 38  7314 38  8328 39  1 90 39  2102
39  3108 39  4111 39  5223 39  6234 39  7239 39  8250 39  9284 39 10328
40  1 25 40  2 69 40  3 82 40  4130 40  5147 40  6184 40  7250 40  8266
40  9267 40 10272 40 11311 40 12325 41  1 45 41  2 55 41  3 63 41  4117
41  5159 41  6186 41  7234 41  8237 41  9252 41 10278 41 11309 41 12311
41 13358 41 14359 42  1 34 42  2 52 42  3152 43  1 35 43  2 51 43  3 52
43  4118 43  5153 43  6177 43  7262 43  8309 44  1  2 44  2 18 44  3 45
44  4 54 44  5 55 44  6 76 44  7 85 44  8108 44  9147 44 10148 44 11166
44 12187 44 13188 44 14234 44 15237 44 16265 44 17330 45  1  4 45  2 70
45  3 75 45  4 77 45  5102 45  6108 45  7122 45  8134 45  9147 45 10150
45 11167 45 12168 45 13221 45 14250 45 15257 45 16262 45 17267 45 18284
45 19294 45 20305 45 21315 45 22327 45 23328 45 24360 45 25364 45 26366
46  1 16 46  2 18 46  3 55 46  4 99 46  5136 46  6223 46  7234 46  8239
46  9312 46 10314 46 11322 46 12327 46 13348 47  1 11 47  2 18 47  3 77
47  4 82 47  5136 47  6137 47  7158 47  8168 47  9178 47 10225 47 11251
47 12290 47 13314 47 14326 47 15347 47 16367 48  1 55 48  2 58 48  3 61
48  4 92 48  5114 48  6118 48  7139 48  8147 48  9148 48 10189 48 11194
48 12201 48 13205 48 14209 48 15210 48 16238 48 17250 48 18255 48 19262
48 20266 48 21267 48 22268 48 23272 48 24278 48 25279 48 26282 48 27294
48 28309 48 29316 48 30328 48 31336 48 32337 48 33339 48 34354 49  1 16
49  2 26 49  3 31 49  4 59 49  5 62 49  6 63 49  7108 49  8139 49  9147
49 10174 49 11185 49 12189 49 13202 49 14204 49 15258 49 16267 49 17290
49 18311 49 19338 49 20344 49 21367 49 22368 50  1136 50  2139 50  3148
50  4152 50  5153 50  6184 50  7237 50  8239 50  9250 50 10261 50 11265
50 12267 50 13284 50 14301 50 15307 50 16312 50 17324 50 18365 51  1  2
51  2 30 51  3 51 51  4 59 51  5 69 51  6102 51  7118 51  8119 51  9307
```

```
51 10342 52  1  4 52  2 27 52  3 35 52  4 37 52  5 39 52  6 42 52  7 43
52  8 44 52  9 60 52 10 68 52 11 69 52 12 70 52 13105 52 14114 52 15126
52 16140 52 17141 52 18154 52 19160 52 20173 52 21207 52 22208 52 23213
52 24221 52 25223 52 26246 52 27252 52 28259 52 29262 52 30272 52 31303
52 32312 52 33313 52 34316 52 35322 52 36342 52 37367 53  1 34 53  2 69
53  3 95 53  4139 53  5240 53  6242 53  7243 53  8299 53  9302 53 10312
53 11319 53 12331 54  1  4 54  2 17 54  3 66 54  4112 54  5150 54  6152
54  7254 54  8265 54  9282 54 10312 54 11324 54 12328 54 13339 54 14340
55  1 54 55  2 70 55  3108 55  4118 55  5122 55  6130 55  7148 55  8150
55  9173 55 10229 55 11242 55 12257 55 13267 55 14268 55 15270 55 16280
55 17303 55 18342 55 19368 56  1  8 56  2 36 56  3 95 56  3147 56  5176
56  6186 56  7246 56  8253 56  9279 56 10284 56 11312 56 12314 56 13343
56 14361 57  1 11 57  2101 57  3168 57  4186 57  5279 57  6312 57  7323
57  8327 58  1  8 58  2 26 58  3 60 58  4 69 58  5168 58  6243 58  7301
58  8314 58  9324 58 10344 59  1 11 59  2 55 59  3 57 59  4 59 59  5 62
59  6 69 59  7 83 59  8118 59  9119 59 10152 59 11170 59 12187 49 13372
59 14314 60  1 11 60  2 99 60  3137 60  4187 60  5221 60  6238 60  7250
60  8272 60  9314 60 10328 61  1 67 61  2 69 61  3104 61  4136 61  5168
61  6190 61  7270 61  8290 61  9327 61 10344 62  1 45 62  2 54 62  3 55
62  4 85 62  5 87 62  6108 62  7114 62  8130 62  9148 62 10183 62 11202
62 12234 62 13305 62 14309 62 15330 62 16361 63  1 30 63  2118 63  3119
63  4148 63  5130 63  6272 63  7365 63  8368 64  1 37 64  2 43 64  3 95
64  4102 64  5126 64  6148 64  7150 64  8153 64  9164 64 10173 64 11184
64 12223 64 13238 64 14259 64 15284 64 16312 64 17331 65  1114 65  2118
65  3150 65  4152 65  5172 65  6173 65  7250 65  8262 65  9292 65 10336
65 11356 65 12359 66  1 85 66  2104 66  3148 66  4150 66  5168 66  6182
66  7316 66  8317 66  9325 66 10328 66 11332 66 12339 66 13342 66 14367
66 15368 67  1 12 67  2 97 66  3104 67  4131 67  5136 67  6159 67  7174
67  8225 67  9280 67 10290 67 11314 67 12326 67 13328 67 14338 67 15343
67 16348 67 17368 68  1 18 68  2 19 68  3 66 68  4 69 68  5 99 68  6104
68  7144 68  8153 68  9166 68 10181 68 11202 68 12243 68 13314 68 14328
68 15348 69  1 25 69  2 55 69  3 73 69  4 83 69  5104 69  6108 69  7116
69  8134 69  9140 69 10148 69 11150 69 12152 69 13158 69 14164 69 15168
69 16262 69 17283 69 18288 69 19317 69 20322 69 21339 69 22342 69 23367
70  1 30 70  2 74 70  3102 70  4108 70  5118 70  6158 70  7184 70  8187
70  9237 70 10251 70 11257 70 12267 70 13290 70 14311 70 15339 70 16364
70 17368 71  1 29 71  2 76 71  3104 71  4108 71  5118 71  6119 71  7162
71  8251 71  9268 71 10331 71 11359 71 12369 72  1 99 72  2 94 72  3112
72  3118 72  4119 72  5194 72  6228 72  7267 72  8270 72  9273 72 10336
73  1 65 73  2 92 73  3108 73  4114 73  5147 73  6148 73  7173 73  8229
73  9257 73 10267 73 11273 73 12284 73 13354 73 14354 74  1 68 74  2102
74  3108 74  4118 74  5148 74  6167 74  7189 74  8267 74  9315 75  1 63
75  2 83 75  3108 75  4113 75  5114 75  6118 75  7162 75  8189 75  9228
75 11250 75 12254 75 13257 75 14268 75 15272 75 16273 75 17284 76  1 25
76  2 58 76  3104 76  4137 76  5148 76  6187 76  7189 76  8265 76  9290
76 10314 76 11325 76 12328 76 13364 76 14368 77  1 11 77  2 69 77  3147
77  4202 77  5229 77  6272 77  7314 77  8316 77  9317 77 10344 78  1152
78  2156 78  3163 78  4203 78  5218 78  6250 78  7254 78  8273 78  9283
78 10284 78 11356 79  1 18 79  2 69 79  3104 79  4123 79  5125 79  6202
79  7217 79  8243 79  9251 79 10290 79 11326 79 12341 79 13344 79 14356
80  1 46 80  2 66 80  3 86 80  4113 80  5151 80  6152 80  7168 80  8192
80  9197 80 10228 80 11240 80 12259 80 13270 80 14273 80•15333 81  1 25
81  2 31 81  3 45 81  4 58 81  5 62 81  6 76 81  7108 81  8140 81  9187
81 10272 81 11290 81 12315 81 13326 81 14348 81 15368 81 16369 82  1 60
82  2 66 82  3128 82  4151 82  5152 82  6303 82  7338 82  8356 83  1 37
83  2 69 83  3147 83  4148 83  5168 83  6181 83  7182 83  8240 83  9317
83 10328 83 11337 84  1 10 84  2 38 84  3 69 84  4 83 84  5179 84  6219
```

```
 84  7243 84  8250 84  9265 84 10273 84 11284 85  1 25 85  2 59 85  3 60
 85  4113 85  5118 85  6189 85  7233 85  8250 85  9257 85 10283 85 11284
 85 12290 85 13311 85 14339 86  1 89 86  2108 86  3118 86  4148 86  5168
 86  6228 85  7233 86  8254 86  9257 86 10267 86 11273 86 12305 86 13354
 86 14364 86 15366 87  1 10 87  2 38 87  3 83 87  4103 87  5178 87  6197
 87  7273 87  8294 87  9338 88  1 11 88  2 21 88  3 31 88  4 69 88  5103
 88  6112 88  7139 88  8151 88  9153 88 10155 88 11218 88 12227 88 13240
 88 14258 88 15270 88 16301 88 17312 88 18326 88 19328 88 20333 88 21341
 89  1 34 89  2 43 89  3 56 89  4 69 89  5116 89  6126 89  7129 89  8131
 89  9153 89 10259 89 11272 89 12291 89 13306 89 14314 90  1 18 90  2 29
 90  3 30 90  4 31 90  5 69 90  6104 90  7108 90  8166 90  9167 90 10180
 90 11240 90 12311 90 13339 90 14369 91  1 18 91  2 22 92  3 68 91  4137
 91  5147 91  6212 91  7237 91  8384 91  9309 91 10314 92  1 30 92  2 69
 92  3102 92  4108 92  5139 92  6148 92  7153 92  8250 92  9268 92 10272
 92 11273 92 12277 92 13284 92 14317 93  1 10 93  2 38 93  3 62 93  4 69
 93  5102 93  6126 93  7127 93  8150 93  9161 93 10178 93 11181 93 12212
 93 13214 93 14250 93 15154 93 16265 93 17284 93 18291 93 19294 93 20365
 94  1 31 94  2 89 94  3 90 94  4108 94  5113 94  6233 94  7254 94  8257
 94  9268 94 10273 94 11356 94 12361 95  1 58 95  2 59 95  3 62 95  4 70
 95  5 85 95  6104 95  7108 95  8140 95  9181 95 10187 95 11240 95 12273
 95 13283 95 14290 95 15311 95 16326 95 17337 95 18347 95 19368 96  1 83
 96  2108 95  3117 96  4118 96  5148 96  6153 96  7228 96  8233 96  9237
 96 10250 96 11257 96 12268 96 13273 96 14283 96 15284 97  1 54 97  2 55
 97  3 83 97  4 89 97  5102 97  6121 97  7122 97  8124 97  9148 97 10166
 97 11175 97 12233 97 13257 97 14275 97 15280 97 16317 97 17328 97 18333
 97 19354 98  1  2 98  2 68 98  3 75 98  4103 98  5118 98  6119 98  7189
 98  8194 98  9197 98 10245 98 11249 98 12336 99  1 45 99  2 55 99  3 70
 99  4134 99  5148 99  6150 99  7174 99  8212 99  9222 99 10265 99 11290
 99 12315 99 13327100  1 25100  2 55100  3 57100  3148100  5167100  6181
100  7187100  8272100  9273100 10310100 11327100 12368101  1 54101  2 69
101  3 77101  4 78101  5 79101  6126101  7148101  8150101  9167101 10240
101 11267101 12284101 13335101 14337101 15339101 16364102  1 69102  2 93
102  3125102  4150102  5153102  6200102  7241102  8250102  9269102 10289
102 11328102 12356102 13365102 14  0102 15  0102 16  0102 17  0102 18  0
```

Appendix D

ILLUSTRATIONS OF COMPUTATIONS TO ESTIMATE RETRIEVAL QUANTITY

Question 1.

Form: $T_1 \cdot T_2 \cdot (T_3 + T_4 + T_5) - T_6$

Term Frequencies: $f(T_1) = 21$

$f(T_2) = 11$

$f(T_3) = 20$

$f(T_4) = 53$

$f(T_5) = 44$

$f(T_6) = 2$

$$f(T_1') = f(T_1 \cdot T_2) = (4.7)\frac{(21 \cdot 11)}{400} = 2.7$$

$$f(T_2') = f(T_3 + T_4) = 20 + 53 - (3.5)\frac{(20 \cdot 53)}{400} = 62$$

$$f(T_3') = f(T_2' + T_5) = 62 + 44 - (3.75)\frac{(62)(44)}{400} = 80$$

$$f(T_4') = f(T_1' \cdot T_3') = \frac{(4)(2.7 \cdot 80)}{400} = 2$$

$$f(T_5') = f(T_4' \cdot T_6) = (100)\frac{(2.2)}{400} = 1$$

$$R_q = \begin{matrix} 2 \text{ IF } f(T_5') = 0 \\ 1 \text{ IF } f(T_5') = 1 \end{matrix}$$

NOTE: All γ's from Fig. 5.43.

Question 8.

Form: $T_1 \cdot (T_2 + T_3)$

Term Frequencies: $f(T_1) = 20$

$f(T_2) = 10$

$f(T_3) = 84$

$$f(T_1{}') = f(T_2 + T_3) = 10 + 84 - (2.8)\frac{(10.84)}{400} \doteq 88$$

$$f(T_2{}') = f(T_1 \cdot T_1{}') = (3.4)\frac{(20 \cdot 88)}{400} = 15$$

$$\therefore \quad R_q = 15$$

Question 14.

Form: $T_1 \cdot T_2$

Term Frequency: $f(T_1) = 38$

$f(T_2) = 31$

$$R_q = (4)\frac{(38 \cdot 31)}{400} \approx 12$$